Immigration, Gender, and Doctoral Dreariness

Aramis D. M. Valverde

12/10/2021

Is It Perhaps Different Doing A PhD In A Country You Grew Up In As Opposed to One That You Moved To? Is Nature Good At Doing Surveys?

Abstract:

A 2019 Nature survey of doctoral students worldwide conducted by Shift Learning, a UK based education market research and consulting firm, demonstrated that PhD students are not doing all that well. The Nature write up by Chris Woolston noted that more than a third of PhD students have sought help for anxiety or depression caused by their their PhD studies, and more than a fifth of PhD students had experienced discrimination or harassment in their PhD program. Prior research has backed up the concept that PhD students, and graduate students at large, aren't doing terribly well. As a PhD student I thought, "yeah, I feel it, its pretty bad, I'm pretty sad.". However I reassured myself that at least I didn't have to deal with being an actual immigrant or a woman. But then one asks oneself, "huh, is it actually worse? Maybe Bill O'Rielly was right when he said that women don't experience discrimination in the workplace.".

To analyze if it is worse to be a PhD student and a woman, or a gender queer/ non binary person, or an immigrant, as opposed to not those things, I analyzed Nature's data set as published on figshare. The analysis took the form of an exploratory data analysis, where many different avenues of discourse were attempted and ideated. I also analyzed the history of the data, its transformations, its chain of custody, and the limitations and problems that the data set may present with.

Overall, a preliminary exploratory data analysis found that gender may have a larger impact on how well a program trains its PhD students than one's status as an immigrant does, that immigrants experience harassment and discrimination more often that non-immigrants, and that there are a good number of concerning issues in the study conducted by Shift Learning for Nature Careers.

Introduction:

Things are hard for PhD students. Research has demonstrated a persistent and severe toll on the mental health of PhD students due to the conditions that they are put under. More than a third have sought help for anxiety and depression caused by their PhD studies, and a fifth of PhD students have experienced discrimination and harassment in their program. A 2020 study of UK PhD students found that 40 percent of said students were at high risk of suicide, with 8 percent having made an attempt.

That being said, every doctoral student's experience is not the same. Some student's experiences are worse. U.S. based doctoral students originating from China also have to deal with isolation, loneliness, and alienation. Some trans students have to deal with severe hostility or insensitivity related to their gender identity or expression, leading suicide attempts rates four times higher than trans students who did not have that severe experience. It is therefore very unfortunate that this survey lumped transgendered and cisgendered people into one variable. Erasure, especially in research, has consequential and damaging effects on trans people.Gender queer and non-conforming was included, so I will attempt an analysis. However, people identifying as such were few, so I may not be able to demonstrate a significant difference even if one is liable to

exist. It is also important to note that gender-specific approaches yield more accurate results, contribute to lessening the paucity of trans research, and are considered best practice.

I must note that I am a minority doctoral student myself. While I am a minority in my department and in my country, I am not an immigrant, nor am I a minority in the state or city where I am studying (Merced, California). I struggle with no language barrier, no one chides me for my voice, and I can travel freely throughout my campus and the city I reside in. I note this because, while my analysis of the data stands on its own, my identity is such that I am not able to speak from direct experience, and that is a limitation within this work.

I was lead to this data set by a tidytuesday post on github, and this research was completed as the final assignment for Dr. Dan Hicks' data science methods course at UC Merced. The data set used in this analysis was sourced from this figshare page. Citations are embedded within the paper and hyperlinked, so you may click on a given word and it will take you to the reference.

Methods

An Analysis of The Data's History

Nature Careers and Shift Insight Nature has had a survey of doctoral students conducted every two years for ten years. The latest survey, conducted in 2019, was advertised by Nature in its "Career News" section, and was titled, "Nature calling: take our PhD survey". This survey was conducted by an English corporation by the name of Shift Insight. Shift Insight is the owning and operating body of Shift Learning, Shift Sustainability, and Shift Membership https://shift-insight.co.uk/about-shift-insight/. This corporation is owned by Jane Powell, a former Publisher at Pearson and Palgrave Macmillan. Shift Insight mainly performs market research, however as with the Nature survey they also conduct public facing research for large institutions. Most staff at Shift Insights have a marketing, writing, or publishing background.

A Faliure to Protect Privacy and Anonymity Their website's language seems to demonstrate commitments to data security, however these amount to either declarations of compliance strictly within the confines of the law, or membership with non-binding commerce and industry associations. Unfortunately for participants, Nature Careers and Shift Insight failed to remove all personally identifying information, specifically the emails of two students and the names of two others, according to my own analysis, where I found two names and emails, along with sensitive and possibly damaging information which was attributable to those persons.

A Non-Tidy Data Set While working with the data I discovered that the excel sheet that was used was not prepared for easy analysis on mediums other than excel itself. In R, the excel sheet holding the data was difficult to work with, requiring the renaming by hand of many columns, the removal of empty columns, the combination of columns that could have been mutually exclusive, and a naming scheme that was less than helpful. Due to the formatting and naming conventions used by the researchers, troubleshooting the importation and use of the data took up far more time than the analysis itself.

The Chain of Command at Shift The finance director and data compliance officer is the person responsible for ensuring that sensitive data is adequately handled. At Shift, that person is Jack Wilson. Jack Wilson appears to have no credentials to handle sensitive data, unlike some of his co-workers, and seems to have been educated at "LAMBDA", without any named degree, according to their linkedin profile. While I could not locate a university by the name of LAMBDA, it is possible that this education was associated with a fraternal organization,and it is also quite possible that Jack Wilson graduated with that education at the top of their class, albeit without a degree. I mention this not just because I want to dunk on these people, but rather because their seemingly irresponsible handling of data could have some real consequences for those people who have had their data exposed. I also mention credentials because Shift appears to have

existed for 20 years but has only had moderately to mildly credentialed persons on staff for seven years. The only visible doctorate holder entered the company on November 2021. This, along with the founders' recent certification and education in R may help explain why the data arrived in a state that was not in line with tidy data principles, why proper attention was not paid to anonymization, and why the analysis of the data was relatively limited. The privacy statement, limited as it was, may have also been broken by the lack of proper anonymization.

The Line of Custody The data appears to have passed through the hands of at least three people, Elsie Lauchlan, the primary researcher for the analysis, David Payne, The managing editor at Careers and Supplements at Nature Springer, and Karen Kaplan, the senior editor of Nature's careers section. According to the version control sheet in the survey script titled "Nature_GradSurvey_Script2019_Final1.docx", there were four versions of the analysis or survey. That is very few given the scope and size of the survey, however it is possible that further analysis could have been forsaken due to cost, as the same paper notes that additional changes may involve extra costs.

Limitations First and foremost I must mention that this data set and the conclusions that have been drawn from it thus far have necessarily been influenced by the funder of the study itself, Nature Careers. While those at Nature Careers may state impartiality or that they do not interfere where there is a conflict of interest, it is important to note that the very focus of the survey itself is guided by those who fund it. If I pay for a survey of cake market share, I may not be pleased when I am delivered instead an analysis of carcinogens in cakes, with my cake factory winning in an evaluation of "most carcinogens per pound-cake". This, is of course not a novel idea. Dr. Federico Germani at University of Zurich, however, took this critique a step further. His analysis and survey demonstrated that most PhD students rate "publishing in high impact journals" as the factor contributing most to their general stress level. The possibility of this analysis downplaying or misrepresenting the publisher's role in PhD student unhappiness is exacerbated not only by the the journal's funding and control over the project, but also by the fact that the survey was advertised by a big, high impact journal, Nature. The connection to the journal Nature was so strong in fact, that some students seemed to erroneously believe that Nature was conducting the survey, further tainting the participant's ability to directly point the finger at publishers at large.

Second is the fact that the data has been cleaned to an extent that has not been disclosed by shift insights. The amount of anonymization and the methods of cleaning the data from the vendor or software that was used to produce the data is not fully and clearly disclosed. Because of this we have no way of discerning what was removed or how the data was manipulated. I also could not attempt the ordinal survey, as the link had expired. The issue of unusual cleaning is exemplified by noting that the "Name" column, which appears to be based on when the individual finished the survey, starts at 1 and ends at 9839. This wouldn't be odd, except that the survey has only 6812 rows. To remove nearly a third of the data set may be justified, however the spacing of these deletions is odd. For instance, until row 349 in the excel sheet, we see no removals have occurred, then at row 350 we have lost our first batch of participants, as the number which has been increasing one by one suddenly jumps up from 347 at line 349 to 356 at 350. Moreover, By line 373 we have jumped all the way to 779. In 24 rows, from 349-373, we have lost 376 individuals! This is not to say that anything suspicious is going on, but rather this is meant to illustrate that we simply cannot properly evaluate what has happened to these values and therefore we cannot evaluate if the loss of those values could have impacted their or our subsequent analysis in any way.

Third is the fact that the participant pool is liable to have been skewed because of the method of advertisement (via Nature Careers), and because Shift Insights seems to reuse participants. This survey was also used to recruit to their participant pool, as evidenced by the question towards the end of the survey asking for contact information and asking if the participant would like to be included in later surveys. Because of their commercial interests and focus on market research, it is possible that their existing participant pool is incidentally selecting against and for some traits. For example, I am not liable to read to an email from Shift Insight, much less participate in their surveys, because I don't respect market researchers. This selection bias may be partially mitigated against by the fact that some PhD students appear to have thought that Nature was conducting the survey, and not Shift Insights. Some students responded as if the people conducting the

survey had some sort of authority, one attempted to report their superiors, another asked for help publishing in Nature.

Fourth is the fact that the survey was only conducted in five languages. Persons from parts of the world who communicate in languages that aren't English, Portuguese, Spanish, Chinese, or French are not likely to have responded to the survey, even if they are competent in the language.

Fifth is a lack of representation. There is little mention or analysis possible for non cis-gendered persons, and there is no capacity for me to make a gender specific analysis of the data as mentioned prior. There is also the issue that persons from certain countries were not given the option to select their country. Taiwan made up the vast majority of write in answers for country of origin in Asia. The point is that overall the survey could do a bit more to provide more specific data, even if in their own analysies they bunch data points up into groups, as with Asia, Australasia, Europe, etc.

Finally, we have the issue that these are self reported data and that that participants may have wanted to introduce their own positive self perceptions. Some participants may have had some concerns about putting this information in a website for fear that their comments may make them identifiable. For at least four students, these sorts of concerns were certainly warranted. The participants may have also wanted to present themselves in a positive light more generally. Given the doctoral community is more than a little neurotic, there are a myriad of manners by which this data could have been skewed by the use of self report measures and not direct measures.

Findings

Immigration

As Defined, Immigration Does Not Have as Large of A Consequence as Gender, But This May Be Explained By An Inadequate Definition of Immigration I found that people who are studying in a place they did not grow up were unlikely to experience differences in training which were large in effect size when compared against people who are studying in the same place as they grew up in. Immigrant students rated their training quality for collecting data, analyzing data, designing experiments, writing for publication, developing resilience, presenting to specialists, presenting in public, finding a satisfying career, managing complex projects, developing a business plan, managing people, managing large budgets, preparedness for non-research science related careers, preparedness for careers that combine academia and industry, and satisfaction with their decision as a phd about the same as non immigrants. However, Immigrants report experiencing fewer hours with their advisors,higher levels of seeking help for anxiety or depression. Immigrants also report to have experienced bullying at slightly higher rates, and discrimination or harassment at significantly higher rates. Immigrants are slightly more likely to be women, however further analysis will be necessary to determine the nature of the relationship between gender, immigration, and outcomes.

The lack of large differences surprised me at first, however I realized that immigration does not necessarily mean that the person is having "the immigrant experience". A PhD student who moved from the U.K. to the U.S. is going to have a different experience than one who moved from China to the U.K.. Gender is also liable to play a role, as is socioeconomic status. The definition of "Are you studying in the same place that you grew up in?" could be re-framed as, "Had you been to the country that you are studying at before you began your collegiate studies?". This analysis is therefore limited due to the lack of specificity within the data set. While it could be possible to derive a more accurate measure, it is beyond the scope of this course.

Gender

Gender Is Important In my analysis, I found that women regularly experience worse outcomes in their education across a span of measures. An analysis demonstrated that, compared to men, women report that their programs have prepared them to carry out scholarly activities less than men. Furthermore, when compared to men, women are less likely to agree or strongly agree to the idea that their program has prepared them for finding careers in industry, academia, and a mix between the two. Gender queer and Non binary

people are, for some measures, more likely to respond having been well prepared by their programs, at times significantly beyond male and female responses. Due to the small sample size of gender queer and non binary people, it is difficult to ascertain whether or not these differences are an artifact of the way the survey was conducted or if there is some other factor accounting for these increased positive measures.

Unfortunately, since gender was the first portion that was completed, I was unable to fully analyze a multitude of measures, particularly for harassment/discrimination, phd satisfaction, and bullying. These will be analyzed at a later date.

Quality of Analysis and Data Set

Problems in the Dataset There appeared to be multiple inconsistencies in timing, particularly the way that they were entered into the final excel sheet. While they appear to be mostly chronologically ordered, there are time stamps which indicate people beginning months after they ended the survey, which isn't possible. This, along with the appearance of some issues with deletions and data cleaning, seem to imply that there was not an adequately ordered mechanism for compiling these survey results. If that is true, there is a high possibility of skewed results, especially when analyzing small subsets of the population. The missing data is also highly clustered, implying multiple avenues of data cleaning and participant removal, which again could skew the results.

My Analysis, Pathways and Conclusions I began by asking myself how immigration plays into being a PhD student. I have known a few international students, and they all outlined a set of struggles, annoyances, and frustrations which were usually above and beyond those I heard from domestic students. I sorted through some tidy tuesday requests, and then found this dataset. Since it was unresolved (in that no one had taken it up), I had to problem solve on my own. Nonetheless, I ensured, through easy to follow analysies, documentation in the form of comments and explanations throughout, and consistent checking against older data frames and outside references, that I hadn't messed up the data or lost anything along the way. I plotted the data for every analysis and problem solved throughout in order to make those plots. Each step is written out and explained to the best of my ability.

That is more than I can say for the data set I inherited and the analysis that sprung from it. The PDF contained in the data folder wasn't bad at all on its own, nor was the analysis that was conducted. My primary issue is with the dataset and the manner in which the survey was conducted and published, as noted in the limitations section within methods. My question at the beginning was, do other genders and people from other places have it worse, and if so, how? I cannot adequately answer those questions in a manner that is not explained better elsewhere. And I certainly cannot come to those conclusions based solely on my analysis of this dataset. Unfortunately I could not analyze the data statistically aside from a cursory set of analysies, and therefore I was not able to adequately quantify the diffrences for gender nor for immigration.

That being said, I was able to evaluate the data set and bring up some concerns about it, and I did set up a strong pipeline for further analysis later on.

##Discussion I found some differences between immigrants and non-immigrants and between the genders as outlined by Nature Careers and Shift Insight. I was not able to fully quantify those differences or conclusively state statistically significant differences. However the pipeline for such an analysis was partially established. In terms of data analysis, I found significant issues with the survey and the dataset created by Shift Insight and Nature. I found personally identifying and potentially damaging information, inconsistent data cleaning and compilation practices, and I cleared up a multitude of issues liable to come up, should another person wish to look at this dataset.

In about two weeks I will close or make private the github repository containing this dataset and my analysis, and then I will seek guidance on how to best address the issue of the personally identifying information I found. It may very well not be my place to do anything about it. I will also continue to clean and adjust this dataset until every variable and value is easily analyzable in and out of R. I did not expect that I would fail to establish significant differences, nor did I expect that the preliminary analysies would indicate that

immigrants and non-immigrants score similarly for satisfaction with their training, as I have heard precisely the opposite. Ultimately this analysis will likely not have an affect on those populations, as I will not publish these non-findings in the condition that they are in. However, I may send a list of recommendations to the editorial folk at Nature Careers and at Shift Insights, and will certainly make something out of this work for publication later on. However only if I obtain actual verifiable results, or alternatively find that those results which have been published are faulty, will I be able to fully evaluate the impact this work will have.

Code and Analysis

In this section I go about creating the graphs and the cursory analysis that is the bulk of this paper.

To hide the whole bit: {r ,results='hide', echo=FALSE,include=FALSE} To run but not output the results: {r echo = T, results = 'hide', message=FALSE, warning=FALSE} Load in Libraries and Installations

```
library(tidyverse)
library(readxl)
library(ggplot2)
library(tinytex)
library(RColorBrewer)
library(ggdist)
library(waffle)
library(dplyr)
```

Here I load in data set that I downloaded from the link above(and which I placed into a "data" folder within the project folder) and check that it imported properly

```
# I downloaded this data from:
# https://figshare.com/articles/dataset/Nature_Graduate_Survey_2017/5480716?file=9558301,
# then placed it into a new folder, christened with the poetic title, "data'.
dfog <- read_excel("data/Nature_PhD survey_Anon_v1.xlsx")
```

Glimpse was used to help ensure that it all was imported properly

```
glimpse(dfog)
```

Head was used to check that the top line had been preserved.

```
head(dfog)
```

The spreadsheet had been loaded in successfully, and it was used to create the data frame, named "dfog", I would analyze and clean. dfog stands for "data frame original".

DATA WRANGLIN' AN' CLEANIN' For Use Later On

Here I created some ancillary data frames, sf1,df2, and df3, that could be used later to ensure that no accidental changes had been made.

```
df1 <- dfog
df2 <- df1
df3 <- df2
```

Here I did some data cleaning by removing the very top row and then using the second row (the questions themselves) as the column's identifiers. The semicolons and periods in the top row created many issues in r, mainly by making the name appear to be a function and making the use of packages like dplyr nearly impossible to use.

```
df4 <- df3
names(df4) <- as.matrix(df4[1, ])
df4 <- df4[-1, ]
df4[] <- lapply(df4, function(x) type.convert(as.character(x)))
```

Here I remove all columns where all values are NA

```
df5 <- df4
df6a <- df5[ , colSums(is.na(df4)) < nrow(df4)]
```

I rename columns that currently have the same name, as that causes issues. I also rename column 117's name in df6a, because the column name in that position is blank. I am not sure how the bottom portion works with attributing df6 to df6a after I declared df6, but it does. I tried it other ways and it didn't output df6. To test this, please do comment out df6 <- df6a and test yourself. I use this work flow to rename all repeated bits, because doing it any other way wouldn't work. Dplyr's tools didn't work for renaming either, and renaming was only accomplished with base r.

```
df6 <- df6a

###AV Note: Column name/question is sourced from Nature_GradSurvey_Script2019_Final1.docx.
df6 <- names(df6a)[117] <- "How much do you expect your PhD to improve your job prospects?"


###AV Rename all duplicates of "If other, please specify" and make unique
df6 <- names(df6a)[12] <- "specified1"
df6 <- names(df6a)[16] <- "specified2"
df6 <- names(df6a)[22] <- "specified3"
df6 <- names(df6a)[34] <- "specified4"
df6 <- names(df6a)[37] <- "specified5"
df6 <- names(df6a)[54] <- "specified6"
df6 <- names(df6a)[79] <- "specified7"
df6 <- names(df6a)[89] <- "specified8"
df6 <- names(df6a)[104] <- "specified9"
df6 <- names(df6a)[116] <- "specified10"
df6 <- names(df6a)[139] <- "specified11"
df6 <- names(df6a)[141] <- "specified12"
df6 <- names(df6a)[146] <- "specified13"
df6 <- names(df6a)[158] <- "specified14"
df6 <- names(df6a)[173] <- "specified15"
df6 <- names(df6a)[183] <- "specified16"
df6 <- names(df6a)[191] <- "specified17"
df6 <- names(df6a)[200] <- "specified18"
df6 <- names(df6a)[225] <- "specified19"
df6 <- names(df6a)[233] <- "specified20"
df6 <- names(df6a)[239] <- "specified21"
df6 <- names(df6a)[255] <- "specified22"
df6 <- names(df6a)[261] <- "specified23"

###AV Rename all duplicates of "What prompted you to study outside your country of
#upbringing" and make unique
df6 <- names(df6a)[23] <- "prompt to study outside country of upbringing1"
df6 <- names(df6a)[24] <- "prompt to study outside country of upbringing2"
df6 <- names(df6a)[25] <- "prompt to study outside country of upbringing3"
df6 <- names(df6a)[26] <- "prompt to study outside country of upbringing4"
df6 <- names(df6a)[27] <- "prompt to study outside country of upbringing5"
df6 <- names(df6a)[28] <- "prompt to study outside country of upbringing6"
df6 <- names(df6a)[29] <- "prompt to study outside country of upbringing7"
df6 <- names(df6a)[30] <- "prompt to study outside country of upbringing8"
df6 <- names(df6a)[31] <- "prompt to study outside country of upbringing9"
```

```
df6 <- names(df6a)[32] <- "prompt to study outside country of upbringing10"
df6 <- names(df6a)[33] <- "prompt to study outside country of upbringing11"

###AV Rename all duplicates of "Who was the perpetrator(s)" and make unique
df6 <- names(df6a)[97] <-  "positionOfPerpetrator1"
df6 <- names(df6a)[98] <-  "positionOfPerpetrator2"
df6 <- names(df6a)[99] <-  "positionOfPerpetrator3"
df6 <- names(df6a)[100] <- "positionOfPerpetrator4"
df6 <- names(df6a)[101] <- "positionOfPerpetrator5"
df6 <- names(df6a)[102] <- "positionOfPerpetrator6"
df6 <- names(df6a)[103] <- "positionOfPerpetrator7"

###AV Rename all duplicates of "Which of the following have you experienced?"
#and make unique
df6 <- names(df6a)[107] <- "experience1"
df6 <- names(df6a)[108] <- "experience2"
df6 <- names(df6a)[109] <- "experience3"
df6 <- names(df6a)[110] <- "experience4"
df6 <- names(df6a)[111] <- "experience5"
df6 <- names(df6a)[112] <- "experience6"
df6 <- names(df6a)[113] <- "experience7"
df6 <- names(df6a)[114] <- "experience8"
df6 <- names(df6a)[115] <- "experience9"

###AV Rename all duplicates of "If you're unlikely to pursue an academic research
#career, what are the main reasons??" and make unique
df6 <- names(df6a)[130] <- "reasonUnlikelyAcademicCarreerPursuit1"
df6 <- names(df6a)[131] <- "reasonUnlikelyAcademicCarreerPursuit2"
df6 <- names(df6a)[132] <- "reasonUnlikelyAcademicCarreerPursuit3"
df6 <- names(df6a)[133] <- "reasonUnlikelyAcademicCarreerPursuit4"
df6 <- names(df6a)[134] <- "reasonUnlikelyAcademicCarreerPursuit5"
df6 <- names(df6a)[135] <- "reasonUnlikelyAcademicCarreerPursuit6"
df6 <- names(df6a)[136] <- "reasonUnlikelyAcademicCarreerPursuit7"
df6 <- names(df6a)[137] <- "reasonUnlikelyAcademicCarreerPursuit8"
df6 <- names(df6a)[138] <- "reasonUnlikelyAcademicCarreerPursuit9"

###AV Rename all duplicates of "How did you arrive at your current career decision?
#Plea..." and make unique
df6 <- names(df6a)[147] <- "HowArriveAtCareerDecision1"
df6 <- names(df6a)[148] <- "HowArriveAtCareerDecision2"
df6 <- names(df6a)[149] <- "HowArriveAtCareerDecision3"
df6 <- names(df6a)[151] <- "HowArriveAtCareerDecision4"
df6 <- names(df6a)[152] <- "HowArriveAtCareerDecision5"
df6 <- names(df6a)[153] <- "HowArriveAtCareerDecision6"
df6 <- names(df6a)[154] <- "HowArriveAtCareerDecision7"
df6 <- names(df6a)[155] <- "HowArriveAtCareerDecision8"
df6 <- names(df6a)[156] <- "HowArriveAtCareerDecision9"
df6 <- names(df6a)[157] <- "HowArriveAtCareerDecision10"

###AV Rename all duplicates of "How do you learn about available career
#opportunities that are beyond academia?" and make unique
df6 <- names(df6a)[159] <- "HowLearnCareerNotAcademia1"
df6 <- names(df6a)[160] <- "HowLearnCareerNotAcademia2"
```

```
df6 <- names(df6a)[161] <- "HowLearnCareerNotAcademia3"
df6 <- names(df6a)[162] <- "HowLearnCareerNotAcademia4"
df6 <- names(df6a)[163] <- "HowLearnCareerNotAcademia5"
df6 <- names(df6a)[164] <- "HowLearnCareerNotAcademia6"
df6 <- names(df6a)[165] <- "HowLearnCareerNotAcademia7"
df6 <- names(df6a)[166] <- "HowLearnCareerNotAcademia8"
df6 <- names(df6a)[167] <- "HowLearnCareerNotAcademia9"
df6 <- names(df6a)[168] <- "HowLearnCareerNotAcademia10"
df6 <- names(df6a)[169] <- "HowLearnCareerNotAcademia11"
df6 <- names(df6a)[170] <- "HowLearnCareerNotAcademia12"
df6 <- names(df6a)[171] <- "HowLearnCareerNotAcademia13"
df6 <- names(df6a)[172] <- "HowLearnCareerNotAcademia14"

###AV Rename all duplicates of "Which of the following 3 things would you say are
#the most difficult for PhD students in your discipline?" and make unique
df6 <- names(df6a)[174] <- "DifficultInDiscipline1"
df6 <- names(df6a)[175] <- "DifficultInDiscipline2"
df6 <- names(df6a)[176] <- "DifficultInDiscipline3"
df6 <- names(df6a)[177] <- "DifficultInDiscipline4"
df6 <- names(df6a)[178] <- "DifficultInDiscipline5"
df6 <- names(df6a)[179] <- "DifficultInDiscipline6"
df6 <- names(df6a)[180] <- "DifficultInDiscipline7"
df6 <- names(df6a)[181] <- "DifficultInDiscipline8"
df6 <- names(df6a)[182] <- "DifficultInDiscipline9"

###AV Rename all duplicates of "Which of the following would you say are the
#most difficult for PhD students in the country where you are studying?" and make unique
df6 <- names(df6a)[184] <- "DifficultInCountry1"
df6 <- names(df6a)[185] <- "DifficultInCountry2"
df6 <- names(df6a)[186] <- "DifficultInCountry3"
df6 <- names(df6a)[187] <- "DifficultInCountry4"
df6 <- names(df6a)[188] <- "DifficultInCountry5"
df6 <- names(df6a)[189] <- "DifficultInCountry6"
df6 <- names(df6a)[190] <- "DifficultInCountry7"

###AV Rename all duplicates of "Which of the following resources do you think PhD
#students need the most in order to establish a satisfying career?" and make unique
df6 <- names(df6a)[192] <- "ResourcesForSatisfyingCareer1"
df6 <- names(df6a)[193] <- "ResourcesForSatisfyingCareer2"
df6 <- names(df6a)[194] <- "ResourcesForSatisfyingCareer3"
df6 <- names(df6a)[195] <- "ResourcesForSatisfyingCareer4"
df6 <- names(df6a)[196] <- "ResourcesForSatisfyingCareer5"
df6 <- names(df6a)[197] <- "ResourcesForSatisfyingCareer6"
df6 <- names(df6a)[198] <- "ResourcesForSatisfyingCareer7"
df6 <- names(df6a)[199] <- "ResourcesForSatisfyingCareer8"

###AV Rename all duplicates of "Which, if any, of the following activities have
#you done to advance your career?" and make unique
df6 <- names(df6a)[217] <- "ActivitiesToAdvanceCareer1"
df6 <- names(df6a)[218] <- "ActivitiesToAdvanceCareer2"
df6 <- names(df6a)[219] <- "ActivitiesToAdvanceCareer3"
df6 <- names(df6a)[220] <- "ActivitiesToAdvanceCareer4"
df6 <- names(df6a)[221] <- "ActivitiesToAdvanceCareer5"
```

```
df6 <- names(df6a)[222] <- "ActivitiesToAdvanceCareer6"
df6 <- names(df6a)[223] <- "ActivitiesToAdvanceCareer7"
df6 <- names(df6a)[224] <- "ActivitiesToAdvanceCareer8"

###AV Rename all duplicates of "Which of the following social media networks have
#you used to build your professional network" and make unique
df6 <- names(df6a)[226] <- "SocialMediaToBuildNetwork1"
df6 <- names(df6a)[227] <- "SocialMediaToBuildNetwork2"
df6 <- names(df6a)[228] <- "SocialMediaToBuildNetwork3"
df6 <- names(df6a)[229] <- "SocialMediaToBuildNetwork4"
df6 <- names(df6a)[230] <- "SocialMediaToBuildNetwork5"
df6 <- names(df6a)[231] <- "SocialMediaToBuildNetwork6"
df6 <- names(df6a)[232] <- "SocialMediaToBuildNetwork7"

###AV Rename all duplicates of "What would you do differently right now if you
#were starting your programme?" and make unique
df6 <- names(df6a)[234] <- "DoDifferently1"
df6 <- names(df6a)[235] <- "DoDifferently2"
df6 <- names(df6a)[236] <- "DoDifferently3"
df6 <- names(df6a)[237] <- "DoDifferently4"
df6 <- names(df6a)[238] <- "DoDifferently5"

###AV Rename all duplicates of "Which of the following best describes you?"
#and make unique
df6 <- names(df6a)[243] <- "BestDescribes1"
df6 <- names(df6a)[244] <- "BestDescribes2"
df6 <- names(df6a)[245] <- "BestDescribes3"
df6 <- names(df6a)[246] <- "BestDescribes4"
df6 <- names(df6a)[247] <- "BestDescribes5"
df6 <- names(df6a)[248] <- "BestDescribes6"
df6 <- names(df6a)[249] <- "BestDescribes7"
df6 <- names(df6a)[250] <- "BestDescribes8"
df6 <- names(df6a)[251] <- "BestDescribes9"
df6 <- names(df6a)[252] <- "BestDescribes10"
df6 <- names(df6a)[253] <- "BestDescribes11"
df6 <- names(df6a)[254] <- "BestDescribes12"

###AV Rename all duplicates of "Do you have any caring responsibilities" and
#make unique
df6 <- names(df6a)[256] <- "CaringResponsibilities1"
df6 <- names(df6a)[257] <- "CaringResponsibilities2"
df6 <- names(df6a)[258] <- "CaringResponsibilities3"
df6 <- names(df6a)[259] <- "CaringResponsibilities4"
df6 <- names(df6a)[260] <- "CaringResponsibilities5"


###AV Final Important Bits, one more rename and the declaration that makes
#renaming work.
df6 <- names(df6a)[242] <- "Gender"
df6 <- names(df6a)[13]  <- "Immigrant"
df6 <- df6a
```

```
###AV Workflow Notes:
###AV This bit told me what ought to be renamed and made not identical
###rename(df6, combine1 = 150)

###AV This is the template I used to fill in the above lines, without
#the "#" infront of each line.

####AV Rename all duplicates of "" and make unique
#df6 <- names(df6a)[] <- ""
#df6 <- names(df6a)[] <- ""
#df6 <- names(df6a)[] <- ""
#df6 <- names(df6a)[] <- ""
#df6 <- names(df6a)[] <- ""


###Finished Product: df7
df7 <- df6
```

Here I separate and then combine demographic data, collapsing the multiple columns that the data is stored in into a single new column, while keeping the original columns, and converting NA values back to NA values from character values "NA", as the unite feature seems to turn NA into "NA" as an indirect consequence of its usage.

```
df8  <- unite(df7, "ethnicityGroupedOther",  243:254, remove = FALSE, na.rm = TRUE)
#df7 %>% unite("ethnicityOtherWriteInIncluded", 2,remove= FALSE) specified22 is variable
#that has the write-in ethnicity data
df9  <- unite(df8, "ethnicityOtherWriteInNotYetIncluded", 244:252, remove = FALSE, na.rm = TRUE)
###AV Ensure that all NA values remain as NA, as the above unite processes seem
#to convert NA to blanks after uniting. Comment out the 2 lines below to confirm.
df10 <- df9
df10[df10 == ""] <- NA
###AV finally combine them all
df11 <- unite(df10, "ethnicityOtherWriteInIncluded",
              c('ethnicityOtherWriteInNotYetIncluded',
                'specified22') , remove = FALSE, na.rm = TRUE)
```

——— Everything Below This Line Probably Requires df12 or Some Variation of df11 ———

I must note that the cleaning of the demographic data will not be used in this analysis, as I am focusing on gender and immigration, and not on ethnicity. I will analyze ethnicity and clean the whole of the dataset at another time, however time has limited my capacity to do everything I want to do here. But this is EDA, so the clear and distinct lack of direction is a feature, not a bug.

Immigration Analysis against Training and Readiness (Q50) Responses and Other Selections

Data Selection, Renaming, And Compillation

Immigration count and Immigrant proportion analysis

```
###Output A Table and dataframe with the values needed
tableImmigrant <- table(df11['Immigrant'])
```

```
ImmigrantCount<- as.data.frame(tableImmigrant)

###Proportional Analysis
ImmigrantProportion <- as.data.frame(table(df11$Immigrant)/length(df11$Immigrant))
###Display Output genderProportion, note that the question asks are you not an immigrant,
#so an immigrant would have answered "No" and a non-immigrant would answer "Yes"

ImmigrantProportion
```

```
##   Var1      Freq
## 1   No 0.3609806
## 2  Yes 0.6390194
```

Creation of different data frames for analysis by Immigrant, against training efficacy for certain tasks(Q50), which will be used in a later analysis.

```
df12Imm <- df11[c(13,55,56,77,78,87,96,106,
                  201,202,203,204,205,206,207,208,209,210,211,212,213,214,215,216,
                  242)]
#df12ImmNo <- df12Imm[df12Imm$'Immigrant' == 'No',]
#df12ImmYes <- df12Imm[df12Imm$'Immigrant' == 'Yes',]
```

Creation of ancillary and working data frames.

```
df13Imm <- df12Imm
df14Imm <- df13Imm
df14Imm <- df13Imm
```

Some data transformations and a rename, so that I could copy and paste from online sources which predominantly use “df” or “data” and not df23 or whatever I’m doing. Here is where I seem to have renamed df12’s Collecting column and not df12m’s, accounting for the difference between the two.

Rename All Columns With Meaningful Names.

```
df15Imm <- names(df12Imm)[1]   <- "Immigration"

df15Imm <- names(df12Imm)[2]   <- "SatisfiedWithDecisionPhD"
df15Imm <- names(df12Imm)[3]   <- "SatisfiedWithPhDExperience"
df15Imm <- names(df12Imm)[4]   <- "HoursAWeekPhDProgram"
df15Imm <- names(df12Imm)[5]   <- "HoursWithSupervisor"
df15Imm <- names(df12Imm)[6]   <- "AnxietyOrDepression"
df15Imm <- names(df12Imm)[7]   <- "Bullying"
df15Imm <- names(df12Imm)[8]   <- "DiscriminationOrHarassment"

df15Imm <- names(df12Imm)[9]   <- "Collecting"
df15Imm <- names(df12Imm)[10]  <- "Analyzing"
df15Imm <- names(df12Imm)[11]  <- "Designing"
df15Imm <- names(df12Imm)[12]  <- "Writing"
df15Imm <- names(df12Imm)[13]  <- "DevResistance"
df15Imm <- names(df12Imm)[14]  <- "PresentingSpecialist"
df15Imm <- names(df12Imm)[15]  <- "PresentingPublic"
df15Imm <- names(df12Imm)[16]  <- "ApplyingFunding"
```

```
df15Imm <- names(df12Imm)[17]  <- "SatisCareer"
df15Imm <- names(df12Imm)[18]  <- "MngCompProj"
df15Imm <- names(df12Imm)[19]  <- "DevBusinessPlan"
df15Imm <- names(df12Imm)[20]  <- "MngPeople"
df15Imm <- names(df12Imm)[21]  <- "MngLargeBudget"
df15Imm <- names(df12Imm)[22]  <- "FeelProgPrepResearch"
df15Imm <- names(df12Imm)[23]  <- "FeelProgPrepScience"
df15Imm <- names(df12Imm)[24]  <- "FeelProgPrepMixIndAcad"

df15Imm <- names(df12Imm)[25]  <- "Gender"



df15Imm <- df12Imm
dfImm <- df15Imm
```

Make values match immigration as opposed to native born distinction. The question was, "Are you studying in the country you grew up in?" Immigrants would answer no to the question, so sine we are calling this immigration we have to account for that issue. I add an intermediary to be safe

```
df12Imm <- df12Imm %>%
    mutate(Immigration = recode(Immigration, Yes = 'FutureNo', No = 'FutureYes' ))
df12Imm <- df12Imm %>%
    mutate(Immigration = recode(Immigration, FutureNo = 'No', FutureYes = 'Yes' ))
dfImm<-df12Imm
```

Analysis for "Collecting data" variable as a proportion of its answers

```
TableCollectingProportion <- table(df12Imm$Collecting)/length(df12Imm$Collecting)


DFCollectingProportion <- as.data.frame(TableCollectingProportion)

DFCollectingProportion
```

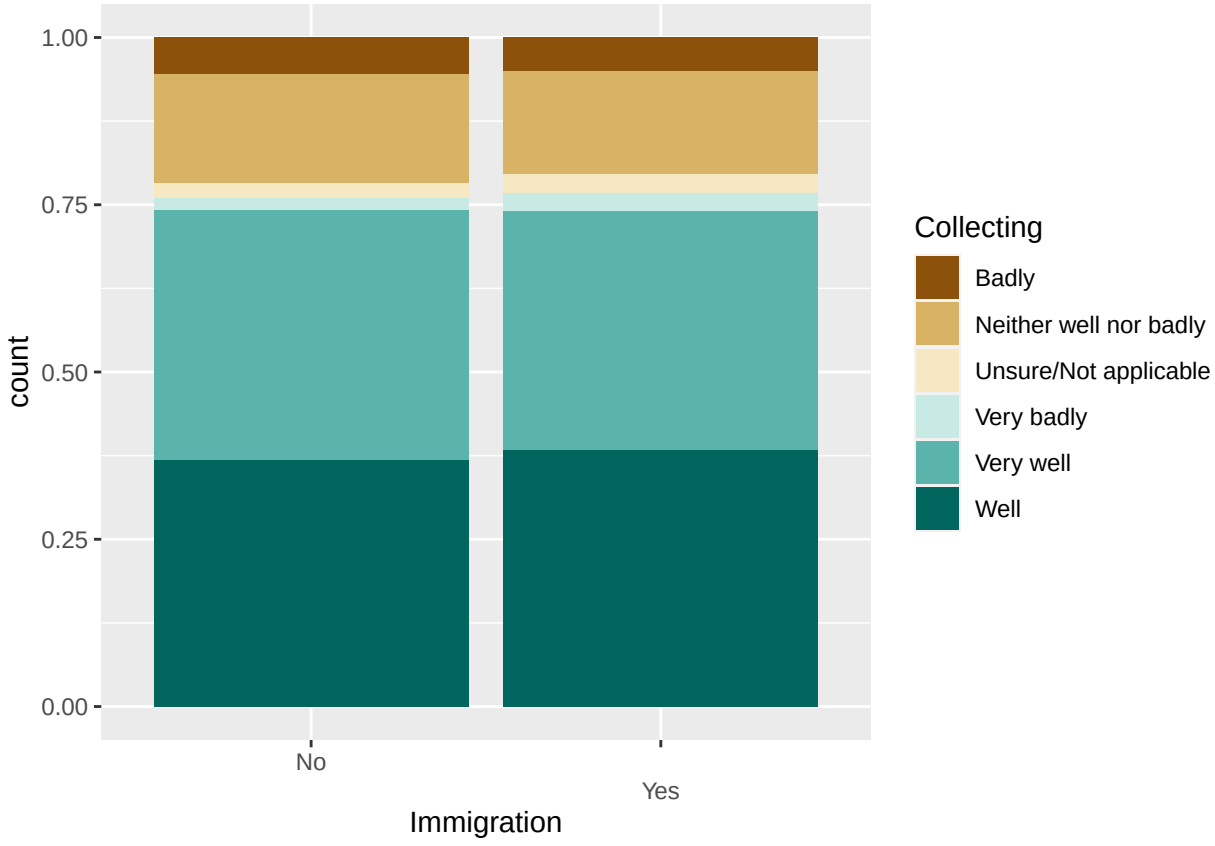
```
##                     Var1       Freq
## 1                 Badly 0.05387551
## 2 Neither well nor badly 0.15942454
## 3  Unsure/Not applicable 0.02422196
## 4             Very badly 0.02172637
## 5              Very well 0.36670581
## 6                   Well 0.37404580
```

Graphs for Immigration

Immigration Against Collecting Column: Collecting Data frame: graphdf1Immigration

```
###New data frame, to ensure there are no issues down the line
graphdf1Immigration <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1Immigration <- graphdf1Immigration[!(graphdf1Immigration$Collecting == "Unsure/Not applicable")
###Refresh old graphdf1Immigration with data_new1 info
graphdf1Immigration <- graphdf1Immigration
###Set variable order
graphdf1Immigration$Collecting <- factor(graphdf1Immigration$Collecting , levels=c("Very badly", "Badly
##Plot
ggplot(data = dfImm) +
  geom_bar(mapping = aes(x = Immigration, fill = Collecting), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

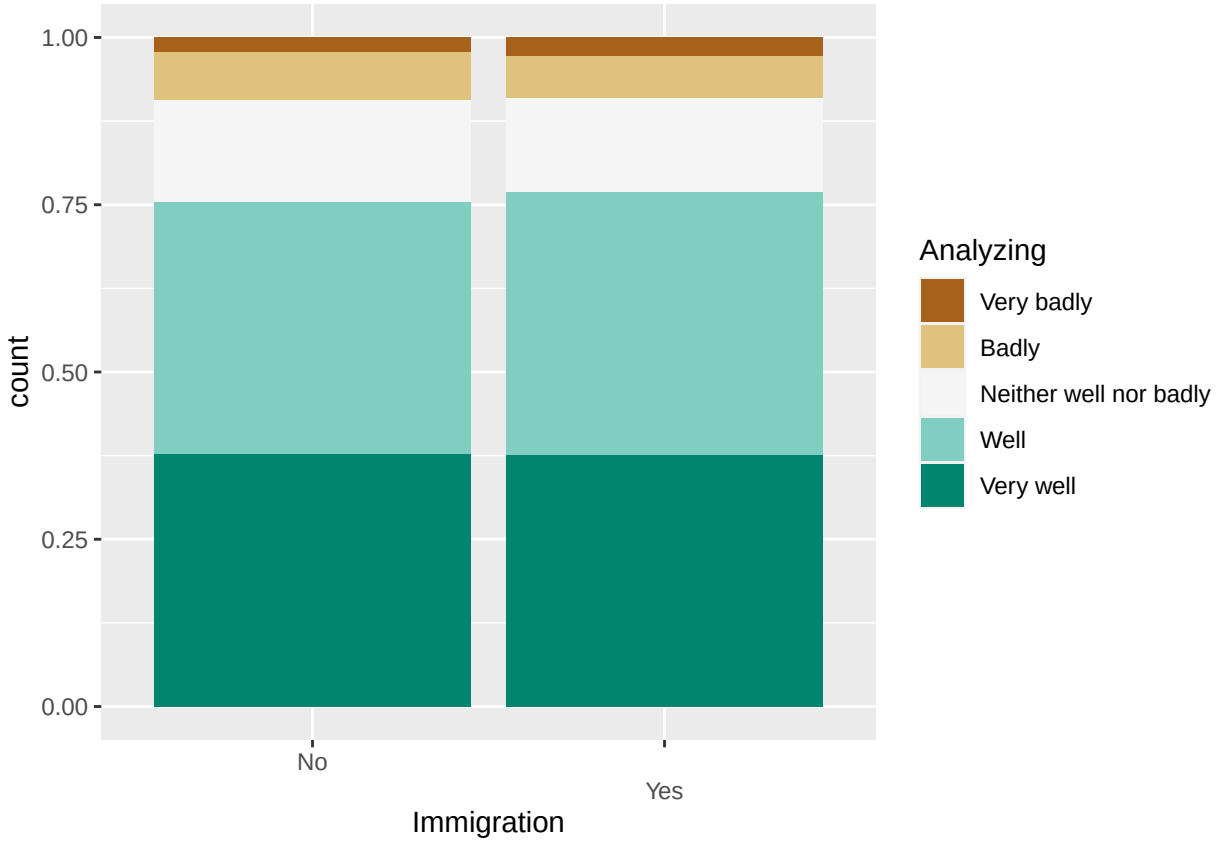


Immigration Against Analyzing Data Column: Analyzing Data frame: graphdfImmigrationAnalyzing

```
graphdfImmigrationAnalyzing <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationAnalyzing<- graphdfImmigrationAnalyzing[!(graphdfImmigrationAnalyzing$Analyzing == "U
# update data frame
graphdfImmigrationAnalyzing <- graphdfImmigrationAnalyzing
#Set variable order
graphdfImmigrationAnalyzing$Analyzing <- factor(graphdfImmigrationAnalyzing$Analyzing , levels=c("Very

#Plot
ggplot(data = graphdfImmigrationAnalyzing) +
  geom_bar(mapping = aes(x = Immigration, fill = Analyzing), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

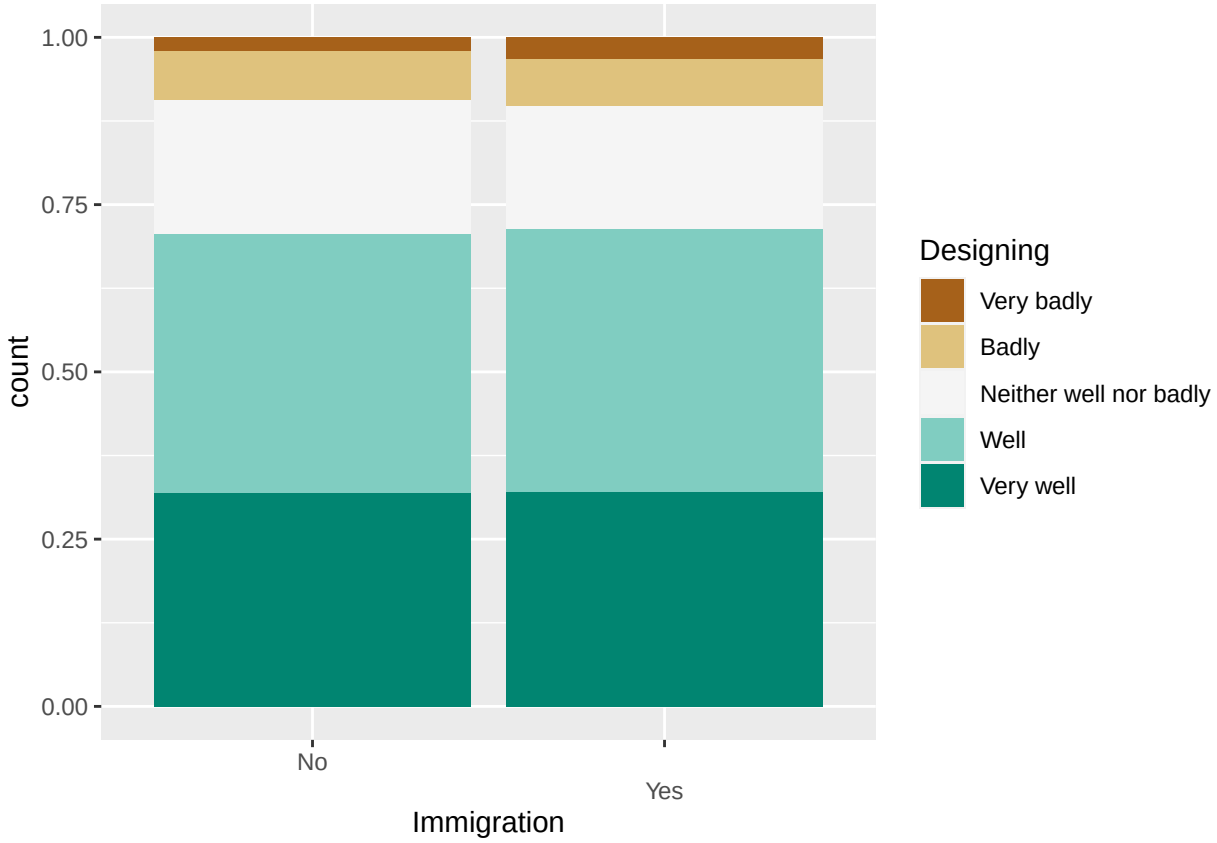


Immigration Against Designing robust reproducible experiments Column: Designing Data frame: graphdfImmigrationDesigning

```
graphdfImmigrationDesigning <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationDesigning<- graphdfImmigrationDesigning[!(graphdfImmigrationDesigning$Designing == "U
# update data frame
graphdfImmigrationDesigning <- graphdfImmigrationDesigning
#Set variable order
graphdfImmigrationDesigning$Designing <- factor(graphdfImmigrationDesigning$Designing , levels=c("Very

#Plot
ggplot(data = graphdfImmigrationDesigning) +
  geom_bar(mapping = aes(x = Immigration, fill = Designing), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

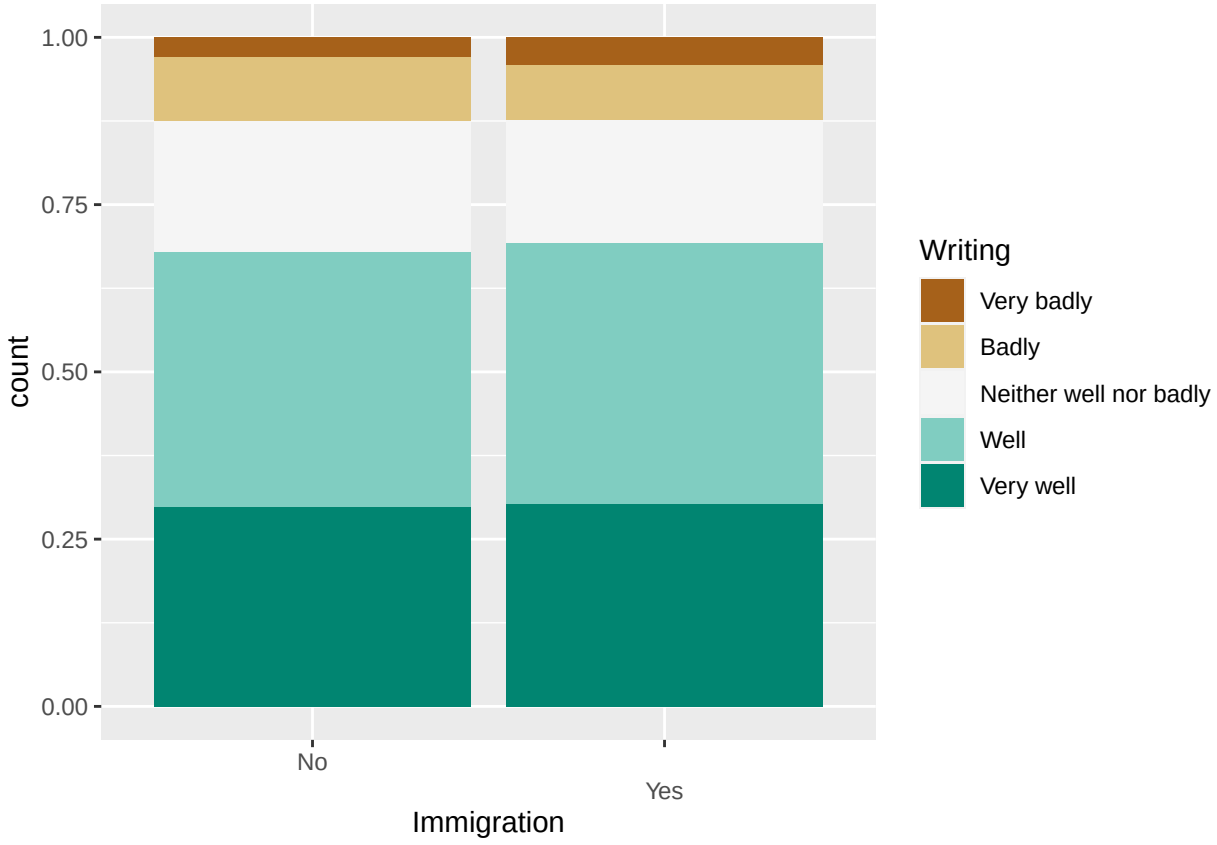


Immigration Against Writing a paper for publication in a peer-reviewed journal Column: Writing graphdfImmigrationWriting

```
graphdfImmigrationWriting <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationWriting<- graphdfImmigrationWriting[!(graphdfImmigrationWriting$Writing == "Unsure/No
# update data frame
graphdfImmigrationWriting <- graphdfImmigrationWriting
#Set variable order
graphdfImmigrationWriting$Writing <- factor(graphdfImmigrationWriting$Writing , levels=c("Very badly",

#Plot
ggplot(data = graphdfImmigrationWriting) +
  geom_bar(mapping = aes(x = Immigration, fill = Writing), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

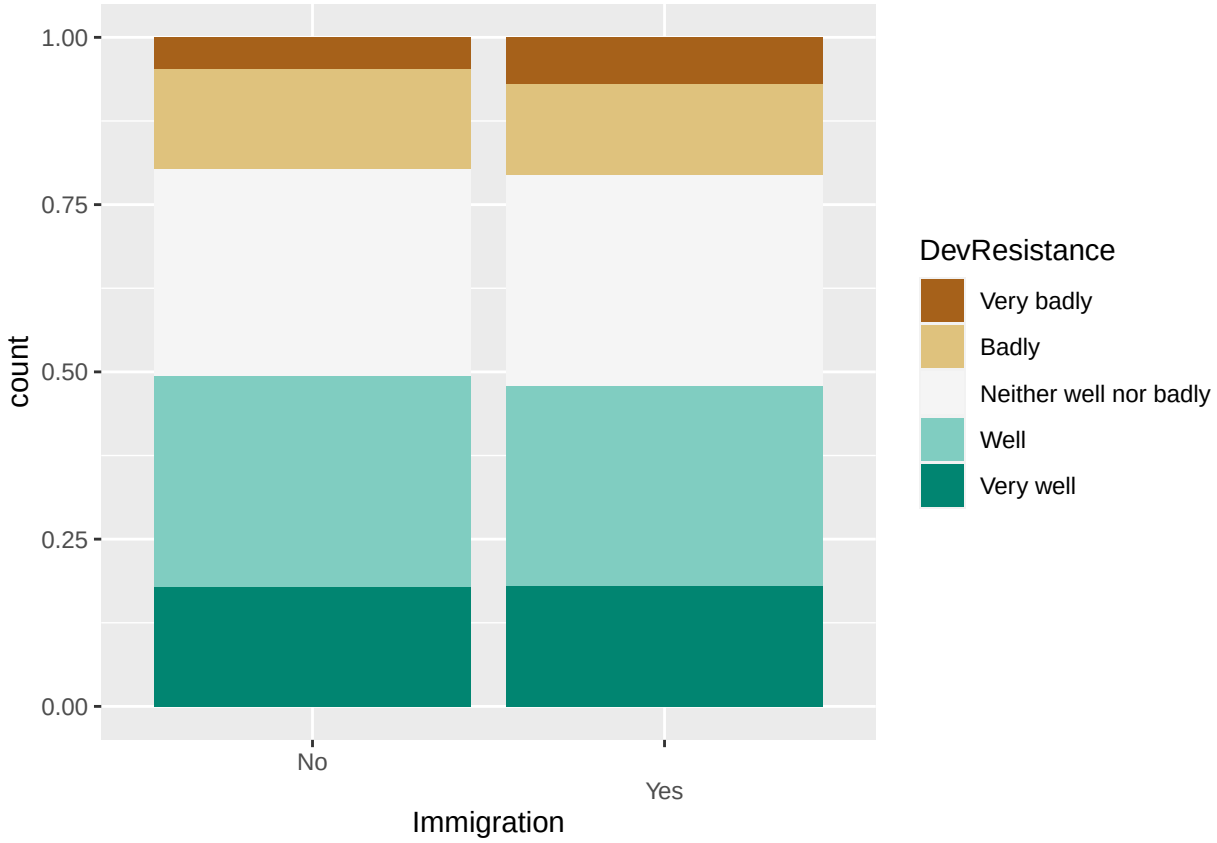


Immigration Against Developing resilience to manage rejection by a peer review panel Column: DevResistance Data frame: graphdfImmigrationDevResistance

```
graphdfImmigrationDevResistance <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationDevResistance<- graphdfImmigrationDevResistance[!(graphdfImmigrationDevResistance$Devl
# update data frame
graphdfImmigrationDevResistance <- graphdfImmigrationDevResistance
#Set variable order
graphdfImmigrationDevResistance$DevResistance <- factor(graphdfImmigrationDevResistance$DevResistance ,

#Plot
ggplot(data = graphdfImmigrationDevResistance) +
  geom_bar(mapping = aes(x = Immigration, fill = DevResistance), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

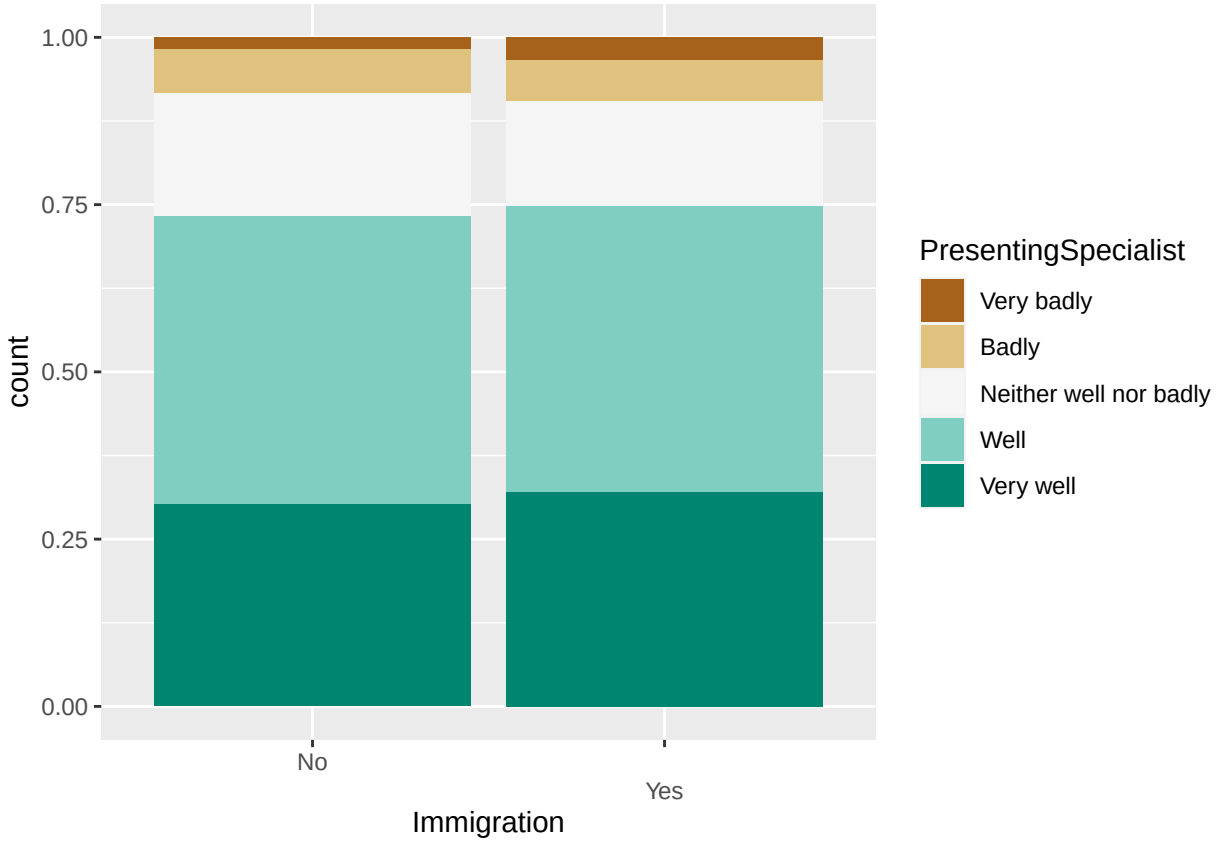


Immigration Against Presenting findings to a specialist audience Column: PresentingSpecialist
Data frame: graphdfImmigrationPresentingSpecialist

```
graphdfImmigrationPresentingSpecialist <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationPresentingSpecialist<- graphdfImmigrationPresentingSpecialist[!(graphdfImmigrationPres
# update data frame
graphdfImmigrationPresentingSpecialist <- graphdfImmigrationPresentingSpecialist
#Set variable order
graphdfImmigrationPresentingSpecialist$PresentingSpecialist <- factor(graphdfImmigrationPresentingSpeci

#Plot
ggplot(data = graphdfImmigrationPresentingSpecialist) +
  geom_bar(mapping = aes(x = Immigration, fill = PresentingSpecialist), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

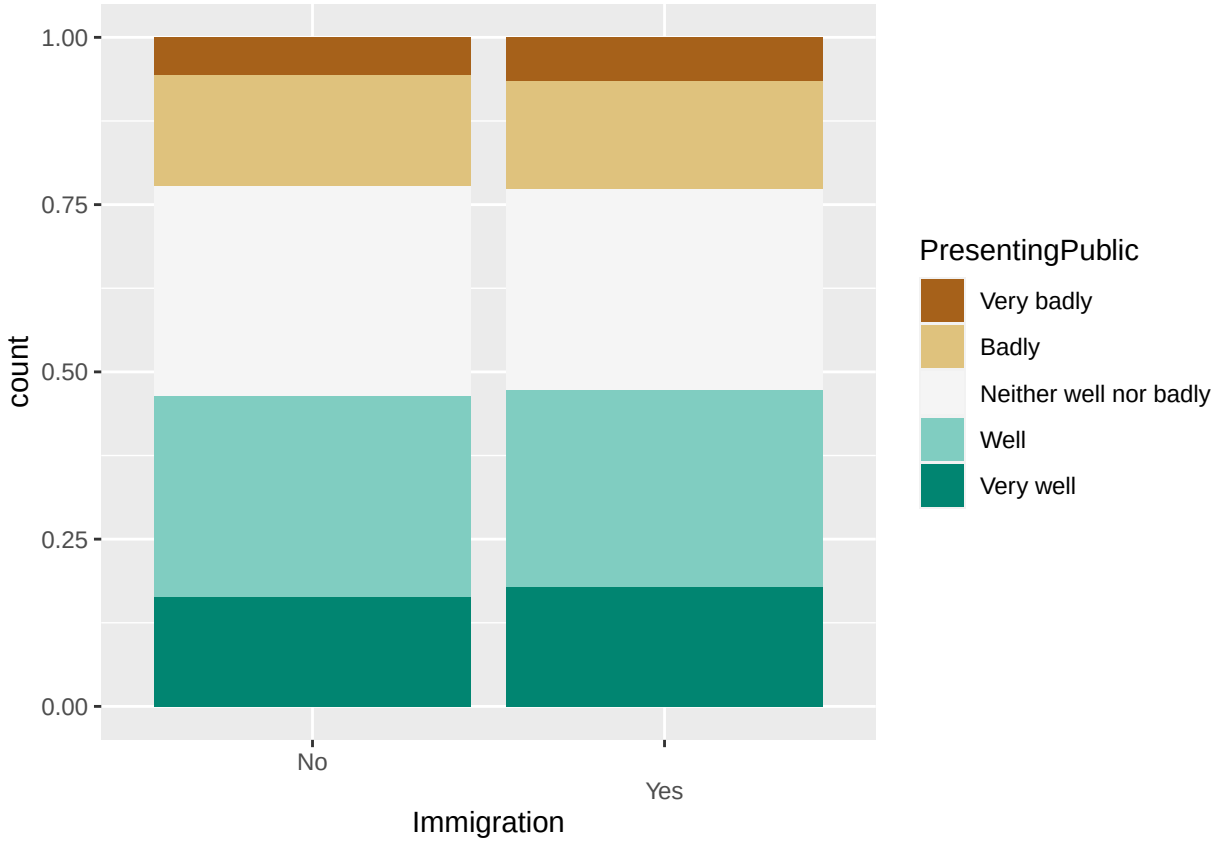


Immigration Against Presenting findings to a non-specialist (public) audience Column: PresentingPublic Data frame: graphdfImmigrationPresentingPublic

```
graphdfImmigrationPresentingPublic <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationPresentingPublic<- graphdfImmigrationPresentingPublic[!(graphdfImmigrationPresentingPu
# update data frame
graphdfImmigrationPresentingPublic <- graphdfImmigrationPresentingPublic
#Set variable order
graphdfImmigrationPresentingPublic$PresentingPublic <- factor(graphdfImmigrationPresentingPublic$Presen

#Plot
ggplot(data = graphdfImmigrationPresentingPublic) +
  geom_bar(mapping = aes(x = Immigration, fill = PresentingPublic), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

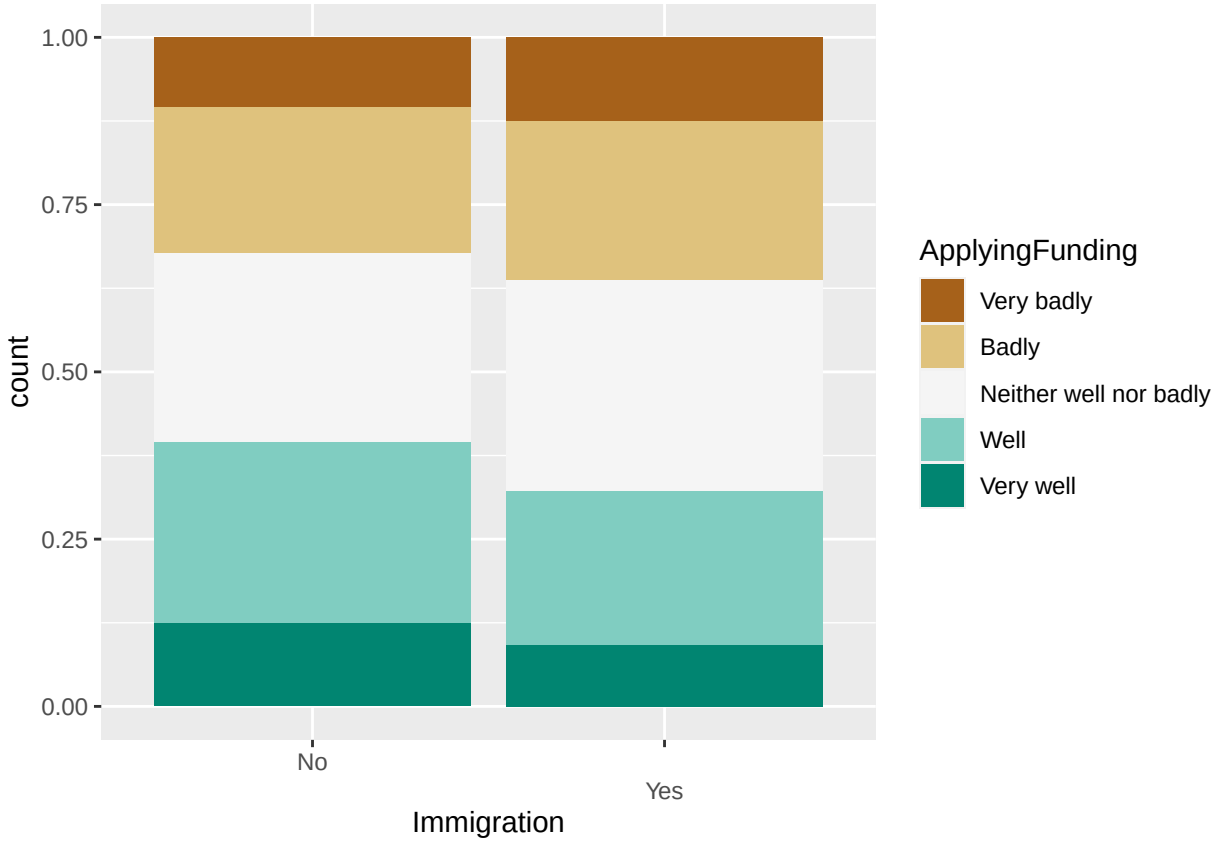


Immigration Against Finding a Applying for funding Column: ApplyingFunding Data frame: graphdfImmigrationApplyingFunding

```
graphdfImmigrationApplyingFunding <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationApplyingFunding<- graphdfImmigrationApplyingFunding[!(graphdfImmigrationApplyingFundi
# update data frame
graphdfImmigrationApplyingFunding <- graphdfImmigrationApplyingFunding
#Set variable order
graphdfImmigrationApplyingFunding$ApplyingFunding <- factor(graphdfImmigrationApplyingFunding$ApplyingFu

#Plot
ggplot(data = graphdfImmigrationApplyingFunding) +
  geom_bar(mapping = aes(x = Immigration, fill = ApplyingFunding), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

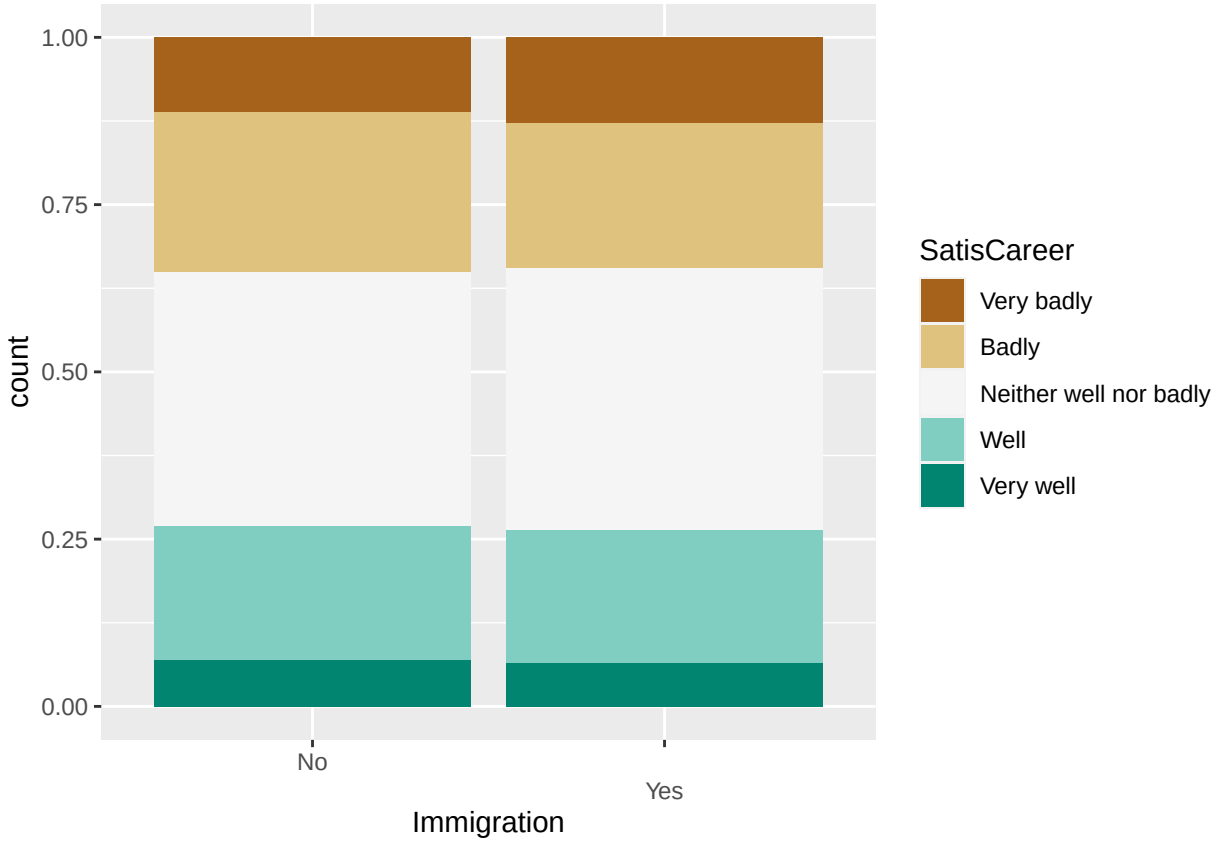


Immigration Against Finding a satisfying career Column: SatisCareer Data frame: graphdfImmigrationSatisCareer

```
graphdfImmigrationSatisCareer <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationSatisCareer<- graphdfImmigrationSatisCareer[!(graphdfImmigrationSatisCareer$SatisCare
# update data frame
graphdfImmigrationSatisCareer <- graphdfImmigrationSatisCareer
#Set variable order
graphdfImmigrationSatisCareer$SatisCareer <- factor(graphdfImmigrationSatisCareer$SatisCareer , levels=

#Plot
ggplot(data = graphdfImmigrationSatisCareer) +
  geom_bar(mapping = aes(x = Immigration, fill = SatisCareer), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

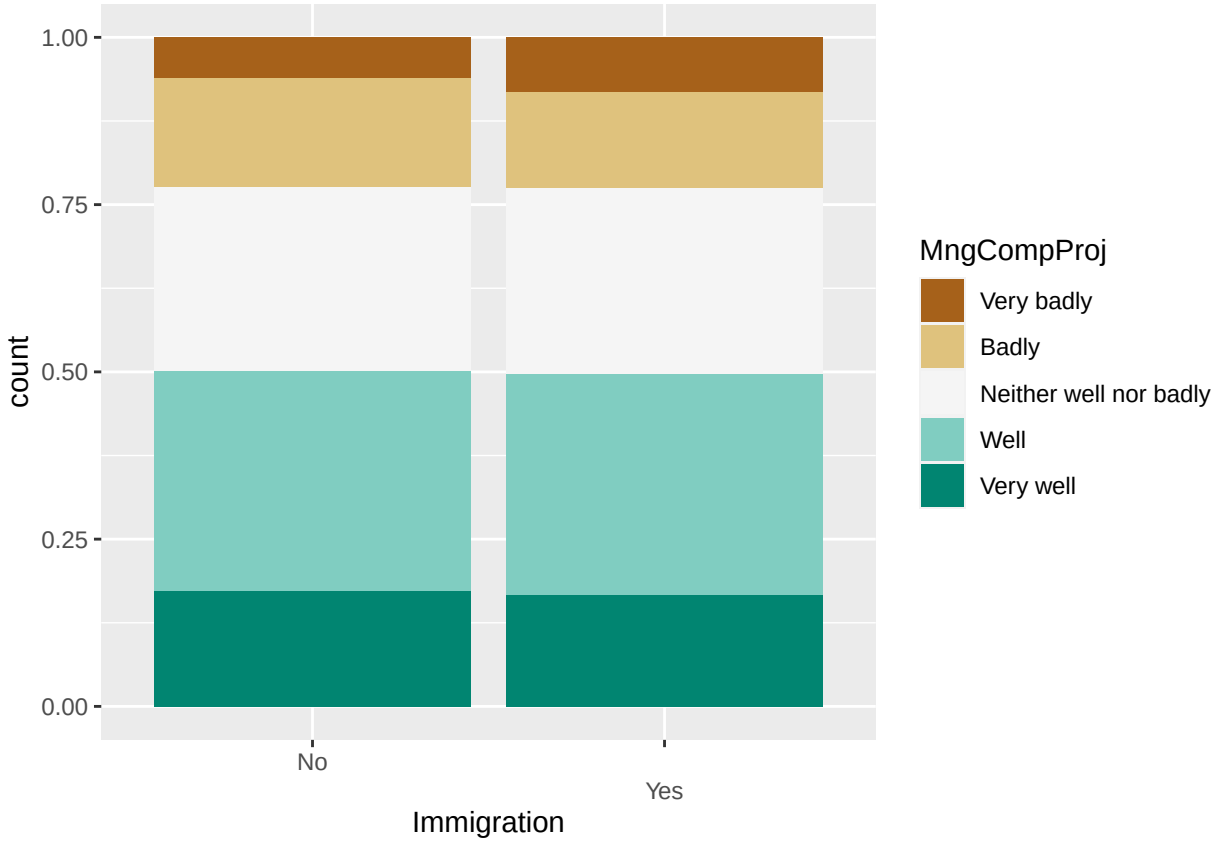


Immigration Against Managing complex projects Column: MngCompProj Data frame: graphdfImmigrationMngCompProj

```
graphdfImmigrationMngCompProj <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationMngCompProj<- graphdfImmigrationMngCompProj[!(graphdfImmigrationMngCompProj$MngCompPr
# update data frame
graphdfImmigrationMngCompProj <- graphdfImmigrationMngCompProj
#Set variable order
graphdfImmigrationMngCompProj$MngCompProj <- factor(graphdfImmigrationMngCompProj$MngCompProj , levels=

#Plot
ggplot(data = graphdfImmigrationMngCompProj) +
  geom_bar(mapping = aes(x = Immigration, fill = MngCompProj), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

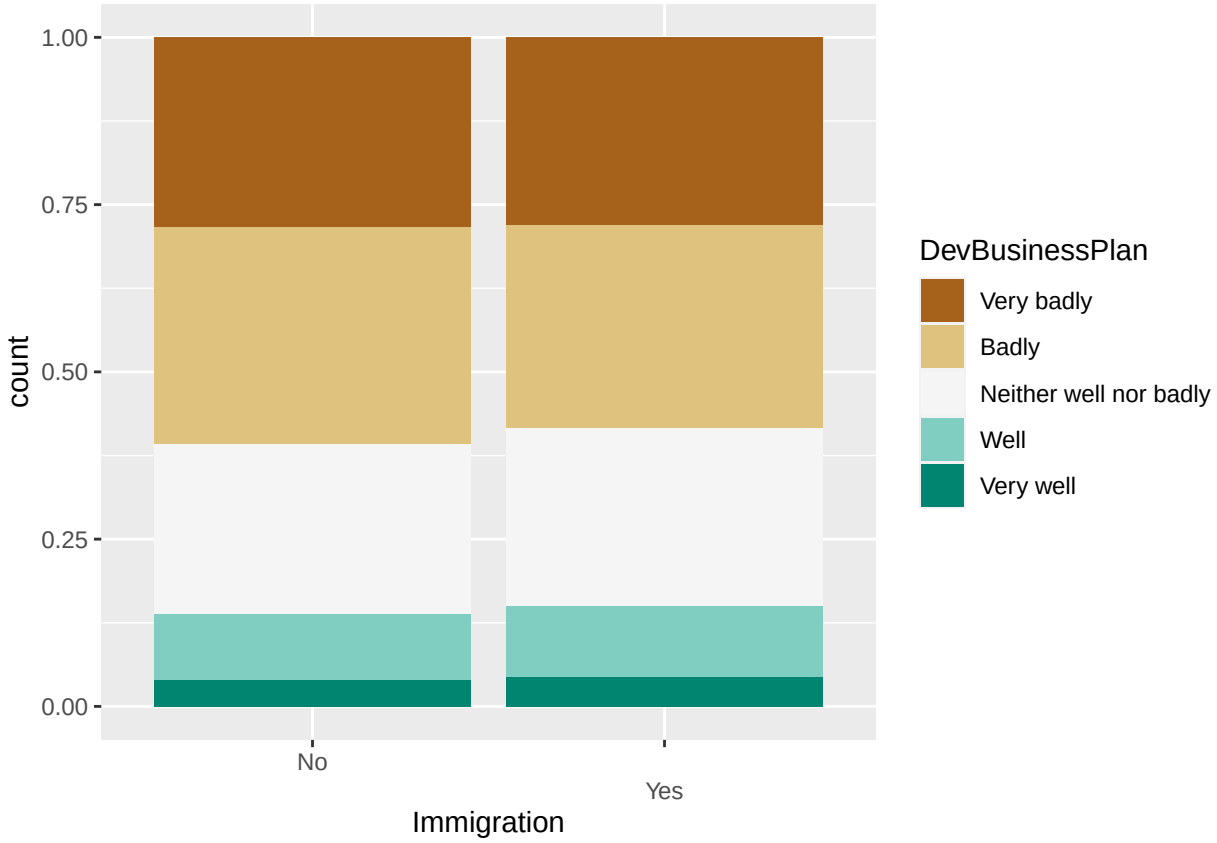


Immigration Against Developing a business plan Column: DevBusinessPlan Data frame: graphdfImmigrationDevBusinessPlan

```
graphdfImmigrationDevBusinessPlan <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationDevBusinessPlan<- graphdfImmigrationDevBusinessPlan[!(graphdfImmigrationDevBusinessPla
# update data frame
graphdfImmigrationDevBusinessPlan <- graphdfImmigrationDevBusinessPlan
#Set variable order
graphdfImmigrationDevBusinessPlan$DevBusinessPlan <- factor(graphdfImmigrationDevBusinessPlan$DevBusines

#Plot
ggplot(data = graphdfImmigrationDevBusinessPlan) +
  geom_bar(mapping = aes(x = Immigration, fill = DevBusinessPlan), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

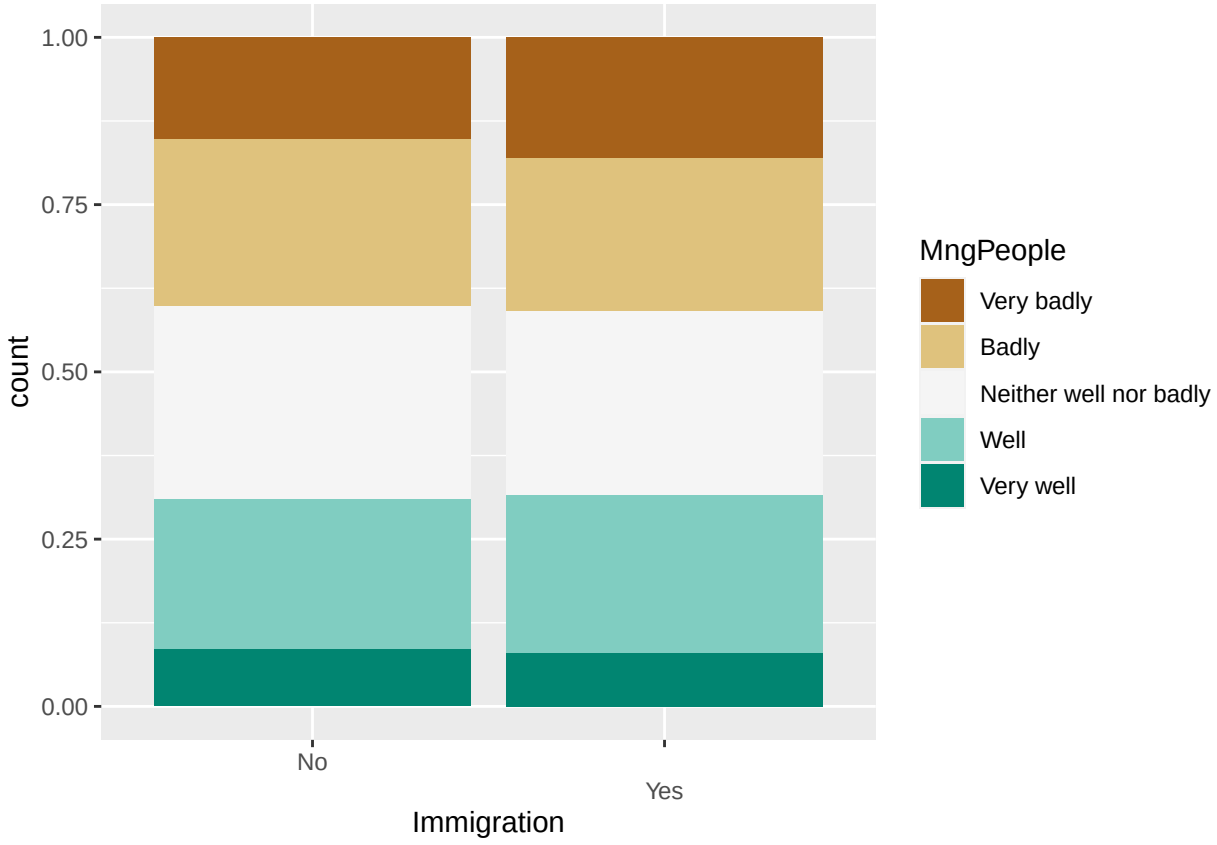


Immigration Against Managing people Column: MngPeople Data frame: graphdfImmigrationMngPeople

```
graphdfImmigrationMngPeople <- dfImm
#Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationMngPeople<- graphdfImmigrationMngPeople[!(graphdfImmigrationMngPeople$MngPeople == "U
# update data frame
graphdfImmigrationMngPeople <- graphdfImmigrationMngPeople
#Set variable order
graphdfImmigrationMngPeople$MngPeople <- factor(graphdfImmigrationMngPeople$MngPeople , levels=c("Very

#Plot
ggplot(data = graphdfImmigrationMngPeople) +
  geom_bar(mapping = aes(x = Immigration, fill = MngPeople), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Immigration Against Managing a large operational budget Column: MngLargeBudget Data frame: graphdfImmigrationMngLargeBudget

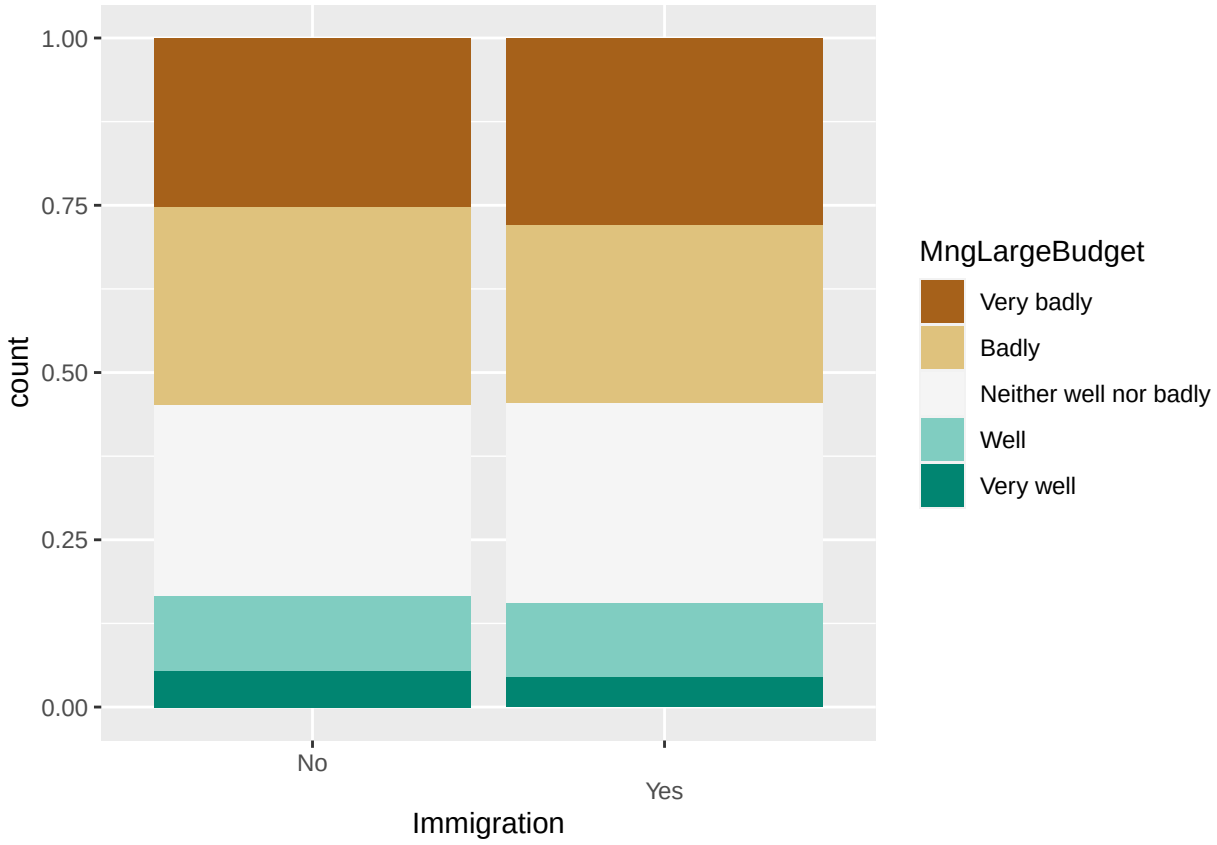
```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationMngLargeBudget <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationMngLargeBudget<- graphdfImmigrationMngLargeBudget[!(graphdfImmigrationMngLargeBudget$

###Refresh old graphdf1gender with data_new1 info
graphdfImmigrationMngLargeBudget <- graphdfImmigrationMngLargeBudget

###Set variable order
graphdfImmigrationMngLargeBudget$MngLargeBudget <- factor(graphdfImmigrationMngLargeBudget$MngLargeBudg

##Plot
ggplot(data = graphdfImmigrationMngLargeBudget) +
  geom_bar(mapping = aes(x = Immigration, fill = MngLargeBudget), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Immigration Against I feel that my programme is preparing me well for a research career
Column: FeelProgPrepResearch Data frame: graphdfImmigrationFeelProgPrepResearch

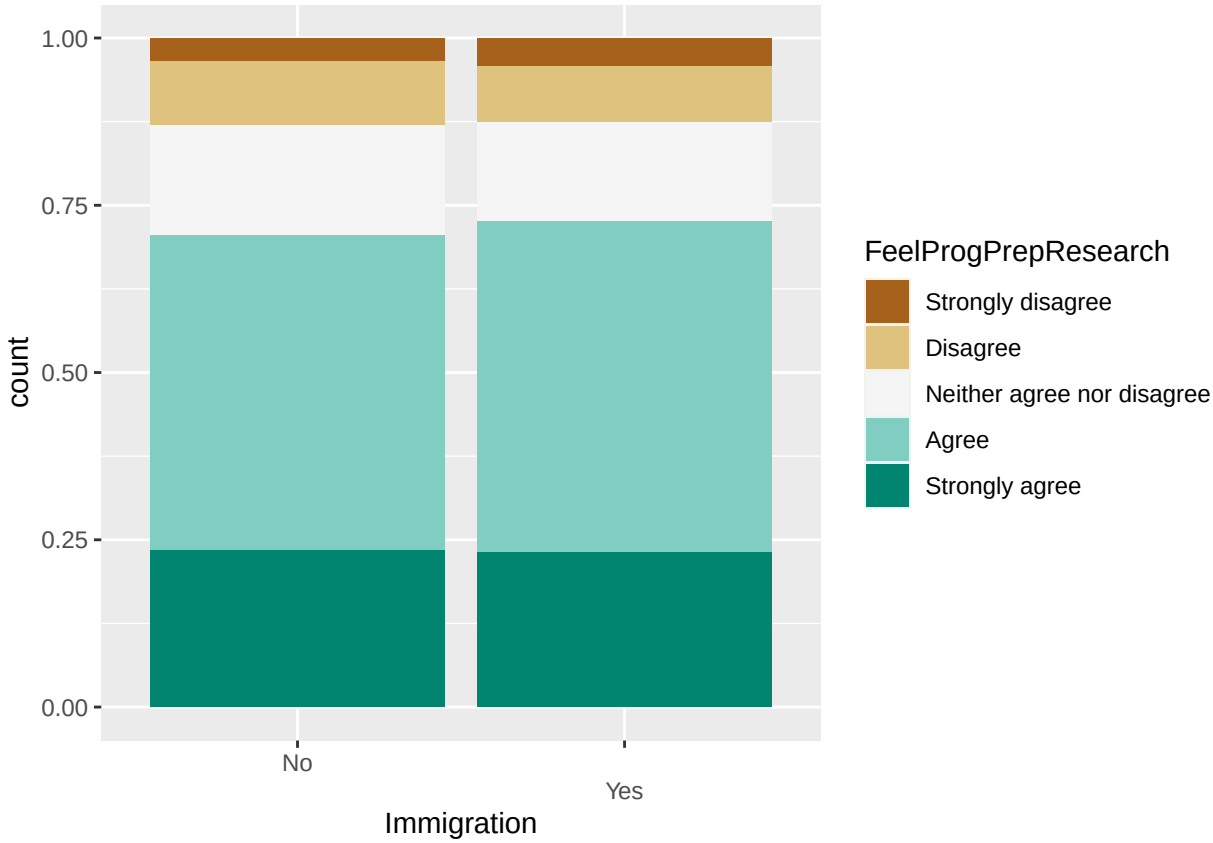
```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationFeelProgPrepResearch <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationFeelProgPrepResearch<- graphdfImmigrationFeelProgPrepResearch[!(graphdfImmigrationFeel

###Refresh old graphdf1gender with data_new1 info
graphdfImmigrationFeelProgPrepResearch <- graphdfImmigrationFeelProgPrepResearch

###Set variable order
graphdfImmigrationFeelProgPrepResearch$FeelProgPrepResearch <- factor(graphdfImmigrationFeelProgPrepRes

##Plot
ggplot(data = graphdfImmigrationFeelProgPrepResearch) +
  geom_bar(mapping = aes(x = Immigration, fill = FeelProgPrepResearch), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Immigration Against I feel that my programme is perparing me well for a non-research science-related career Column: FeelProgPrepScience Data frame: graphdfImmigrationFeelProgPrepScience

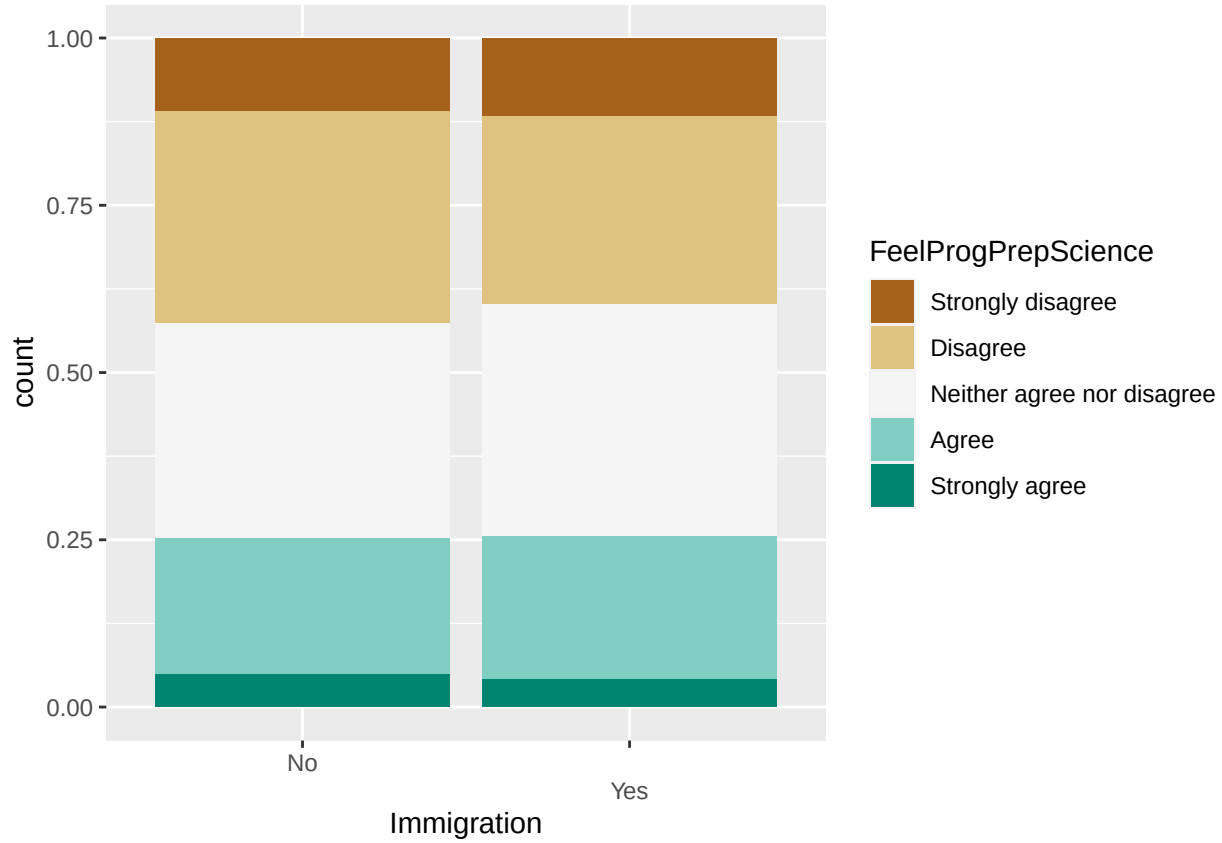
```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationFeelProgPrepScience <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationFeelProgPrepScience<- graphdfImmigrationFeelProgPrepScience[!(graphdfImmigrationFeelPr

###Refresh old graphdf1gender with data_new1 info
graphdfImmigrationFeelProgPrepScience <- graphdfImmigrationFeelProgPrepScience

###Set variable order
graphdfImmigrationFeelProgPrepScience$FeelProgPrepScience <- factor(graphdfImmigrationFeelProgPrepScien

##Plot
ggplot(data = graphdfImmigrationFeelProgPrepScience) +
  geom_bar(mapping = aes(x = Immigration, fill = FeelProgPrepScience), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Immigration Against I feel that my programme is preparing me well for a career that straddles both industry and academia Column: FeelProgPrepMixIndAcad Data frame: graphdfImmigrationFeelProgPrepMixIndAcad

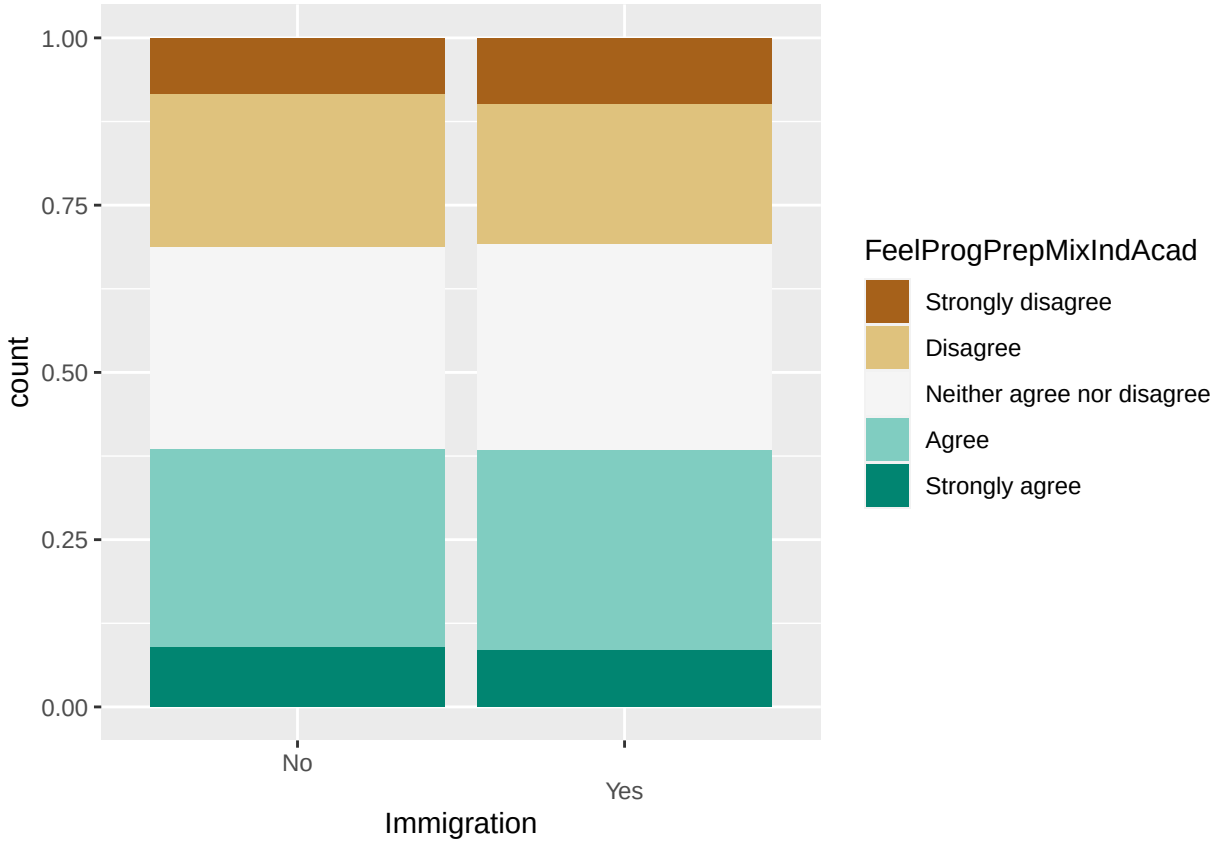
```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationFeelProgPrepMixIndAcad <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationFeelProgPrepMixIndAcad<- graphdfImmigrationFeelProgPrepMixIndAcad[!(graphdfImmigratio

###Refresh old graphdf1gender with data_new1 info
graphdfImmigrationFeelProgPrepMixIndAcad <- graphdfImmigrationFeelProgPrepMixIndAcad

###Set variable order
graphdfImmigrationFeelProgPrepMixIndAcad$FeelProgPrepMixIndAcad <- factor(graphdfImmigrationFeelProgPrep

##Plot
ggplot(data = graphdfImmigrationFeelProgPrepMixIndAcad) +
  geom_bar(mapping = aes(x = Immigration, fill = FeelProgPrepMixIndAcad), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



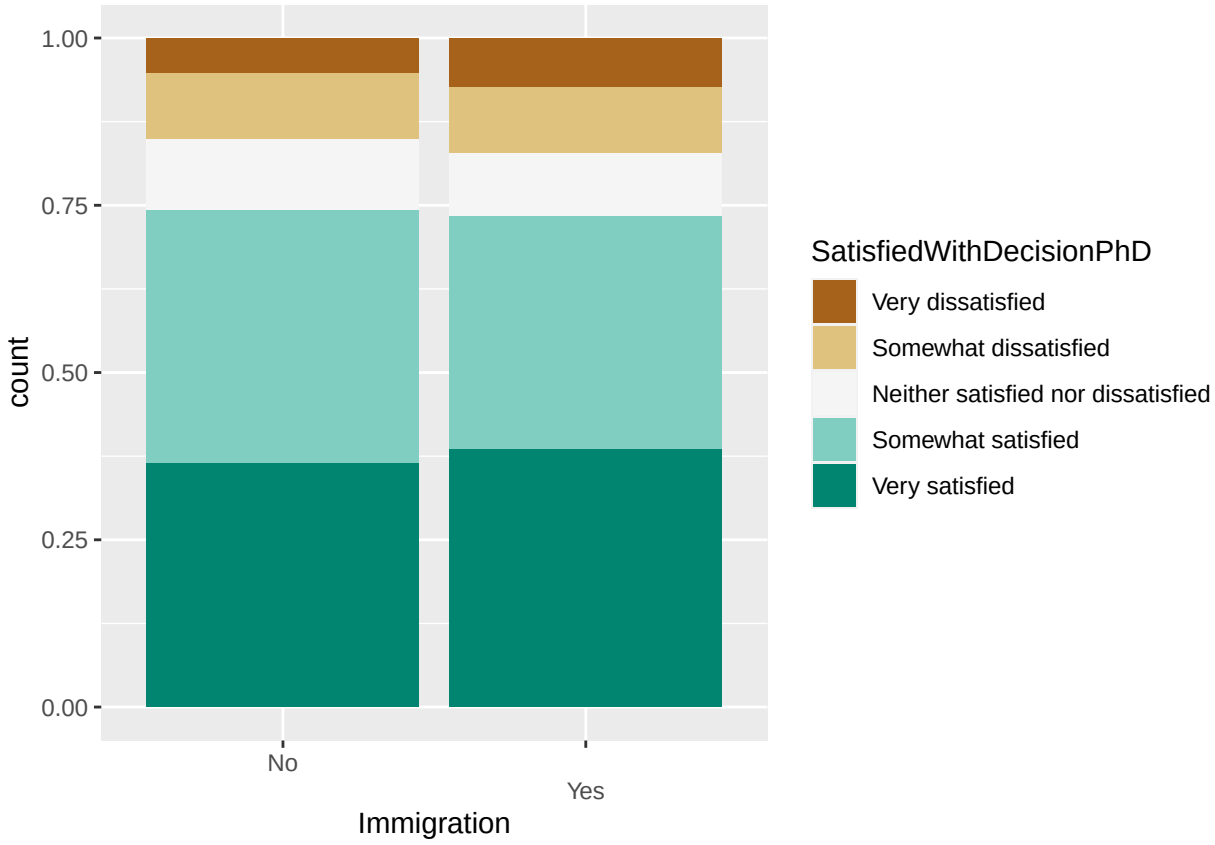
Immigration Against I am satisfied with my decision to pursue a PhD Column: SatisfiedWithDecisionPhD Data frame: graphdfImmigrationSatisfiedWithDecisionPhD

```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationSatisfiedWithDecisionPhD <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
#graphdfImmigrationSatisfiedWithPhDExperience<- graphdfImmigrationSatisfiedWithPhDExperience[!(graphdfI

###Refresh old graphdf1gender with data_new1 info
#graphdfImmigrationSatisfiedWithPhDExperience <- graphdfImmigrationSatisfiedWithPhDExperience
#
###Set variable order
graphdfImmigrationSatisfiedWithDecisionPhD$SatisfiedWithDecisionPhD <- factor(graphdfImmigrationSatisfi

##Plot
ggplot(data = graphdfImmigrationSatisfiedWithDecisionPhD) +
  geom_bar(mapping = aes(x = Immigration, fill = SatisfiedWithDecisionPhD), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Immigration Against I am satisfied with my experience Column: SatisfiedWithPhDExperience
Data frame: graphdfImmigrationSatisfiedWithPhDExperience

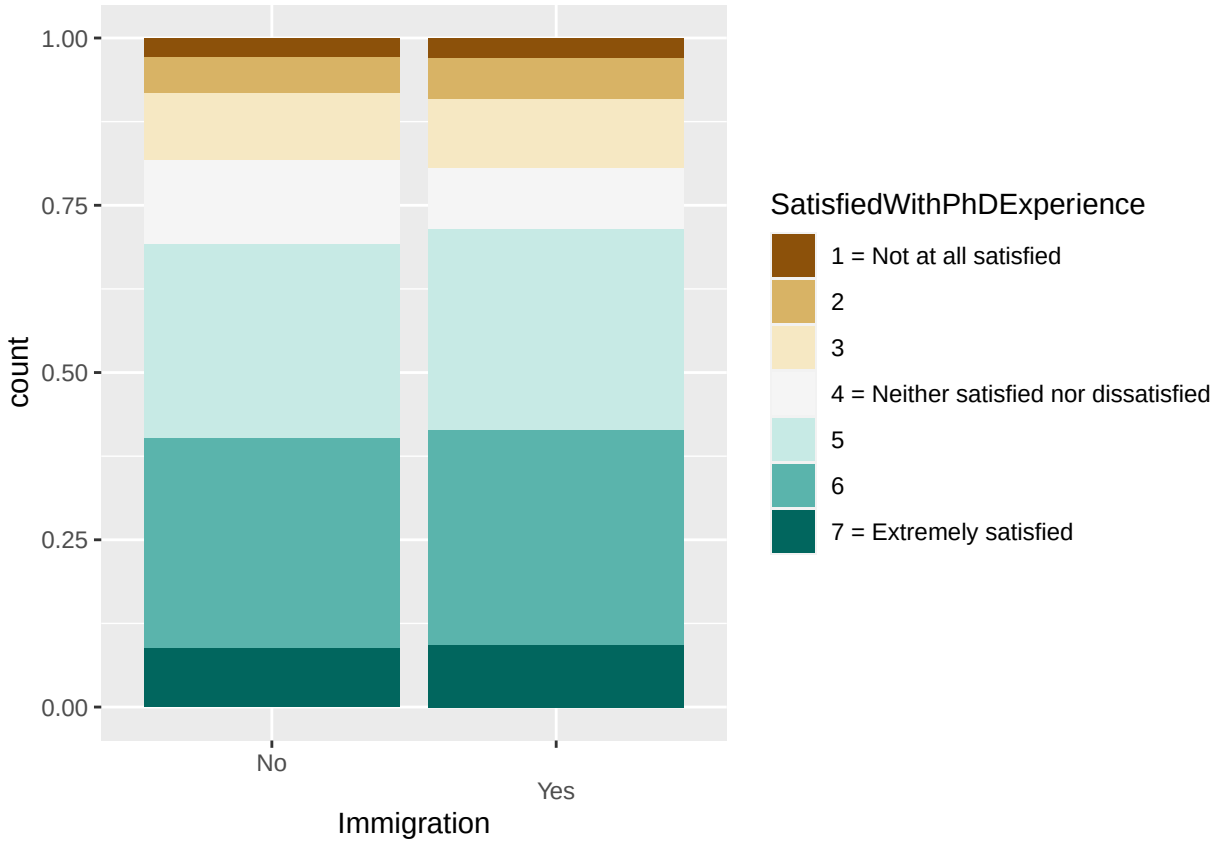
```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationSatisfiedWithPhDExperience <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
#graphdfImmigrationSatisfiedWithPhDExperience<- graphdfImmigrationSatisfiedWithPhDExperience[!(graphdfI

###Refresh old graphdf1gender with data_new1 info
graphdfImmigrationSatisfiedWithPhDExperience <- graphdfImmigrationSatisfiedWithPhDExperience

###Set variable order
graphdfImmigrationSatisfiedWithPhDExperience$SatisfiedWithPhDExperience <- factor(graphdfImmigrationSati

##Plot
ggplot(data = graphdfImmigrationSatisfiedWithPhDExperience) +
  geom_bar(mapping = aes(x = Immigration, fill = SatisfiedWithPhDExperience), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Immigration Against Hours a week on my program Column: HoursAWeekPhDProgram Data frame: graphdfImmigrationHoursAWeekPhDProgram

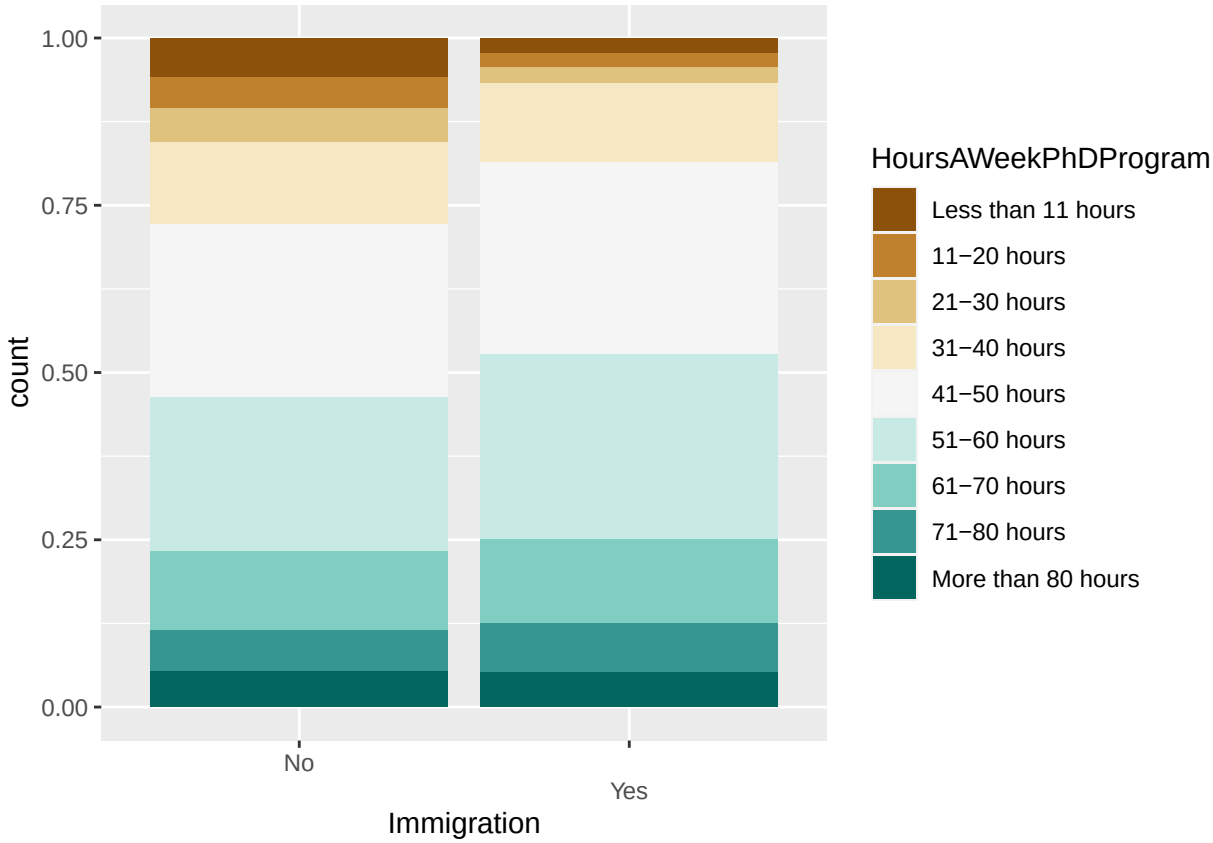
```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationHoursAWeekPhDProgram <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
#graphdfImmigrationHoursAWeekPhDProgram<- graphdfImmigrationHoursAWeekPhDProgram[!(graphdfImmigrationHo

###Refresh old graphdf1gender with data_new1 info
#graphdfImmigrationHoursAWeekPhDProgram <- graphdfImmigrationHoursAWeekPhDProgram

###Set variable order
graphdfImmigrationHoursAWeekPhDProgram$HoursAWeekPhDProgram <- factor(graphdfImmigrationHoursAWeekPhDPr

##Plot
ggplot(data = graphdfImmigrationHoursAWeekPhDProgram) +
  geom_bar(mapping = aes(x = Immigration, fill = HoursAWeekPhDProgram), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Immigration Against Hours a week with my advisor Column: HoursWithSupervisor Data frame: graphdfImmigrationHoursWithSupervisor

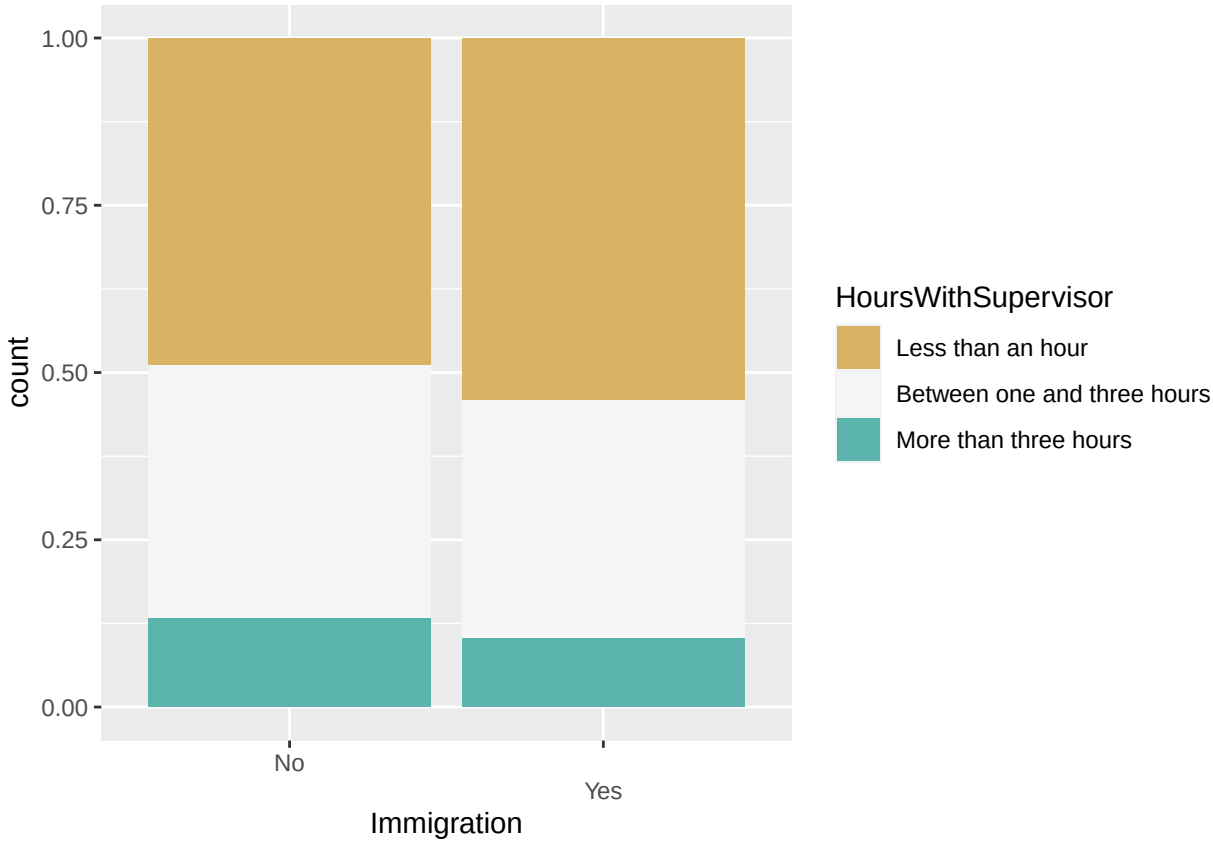
```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationHoursWithSupervisor <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
graphdfImmigrationHoursWithSupervisor <- graphdfImmigrationHoursWithSupervisor[!(graphdfImmigrationHours

###Refresh old graphdf1gender with data_new1 info
#graphdfImmigrationHoursWithSupervisor <- graphdfImmigrationHoursWithSupervisor

###Set variable order
graphdfImmigrationHoursWithSupervisor$HoursWithSupervisor <- factor(graphdfImmigrationHoursWithSupervis

##Plot
ggplot(data = graphdfImmigrationHoursWithSupervisor) +
  geom_bar(mapping = aes(x = Immigration, fill = HoursWithSupervisor), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



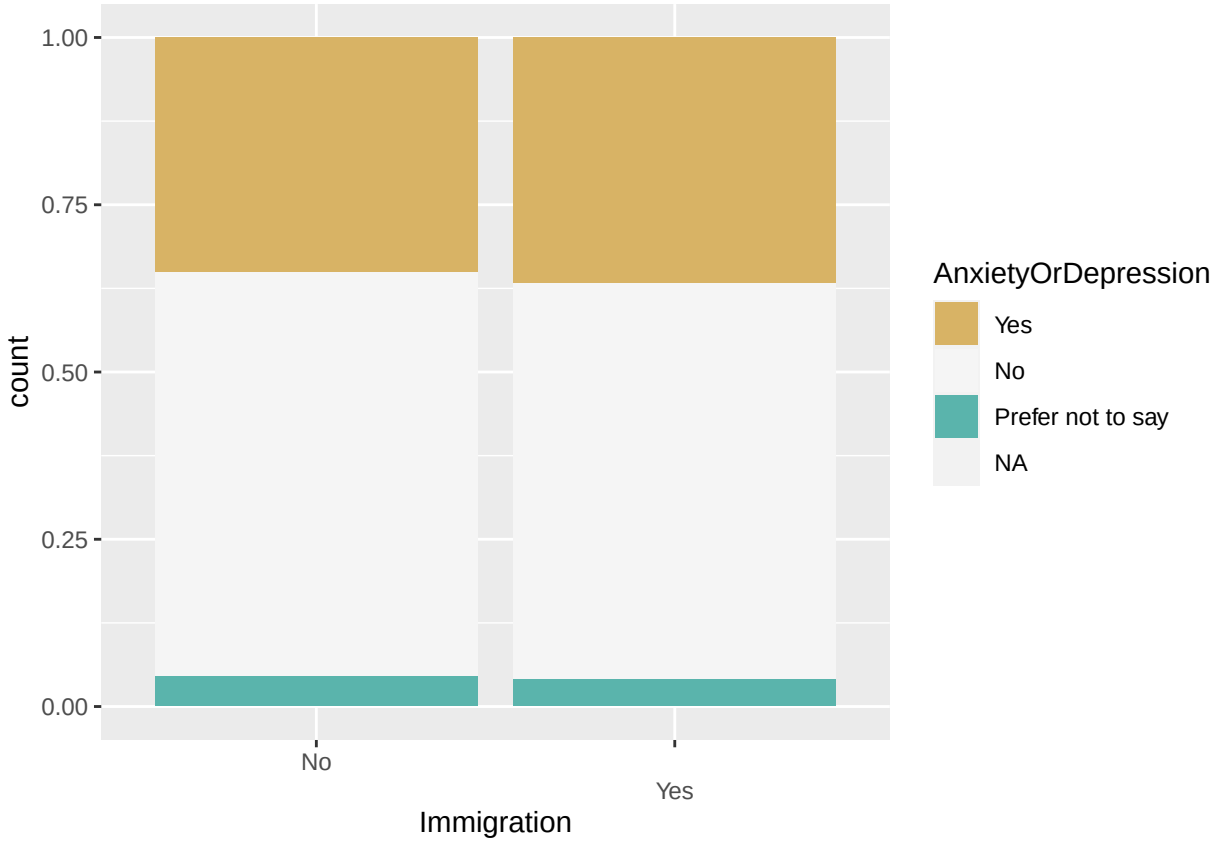
Immigration Against I have seeked help for phd anxiety or deppression Column: AnxietyOrDepression Data frame: graphdfImmigrationAnxietyOrDepression Note: (Kept NA values since they could be interpreted as apathy or negativity, however inclusion doesnt seem to add mcuh to the overall image)

```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationAnxietyOrDepression <- dfImm

###Set variable order
graphdfImmigrationAnxietyOrDepression$AnxietyOrDepression <- factor(graphdfImmigrationAnxietyOrDepressio

##Plot
ggplot(data = graphdfImmigrationAnxietyOrDepression) +
  geom_bar(mapping = aes(x = Immigration, fill = AnxietyOrDepression), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Immigration Against Bullying Experienced Column: Bullying Data frame: graphdfImmigrationBullying

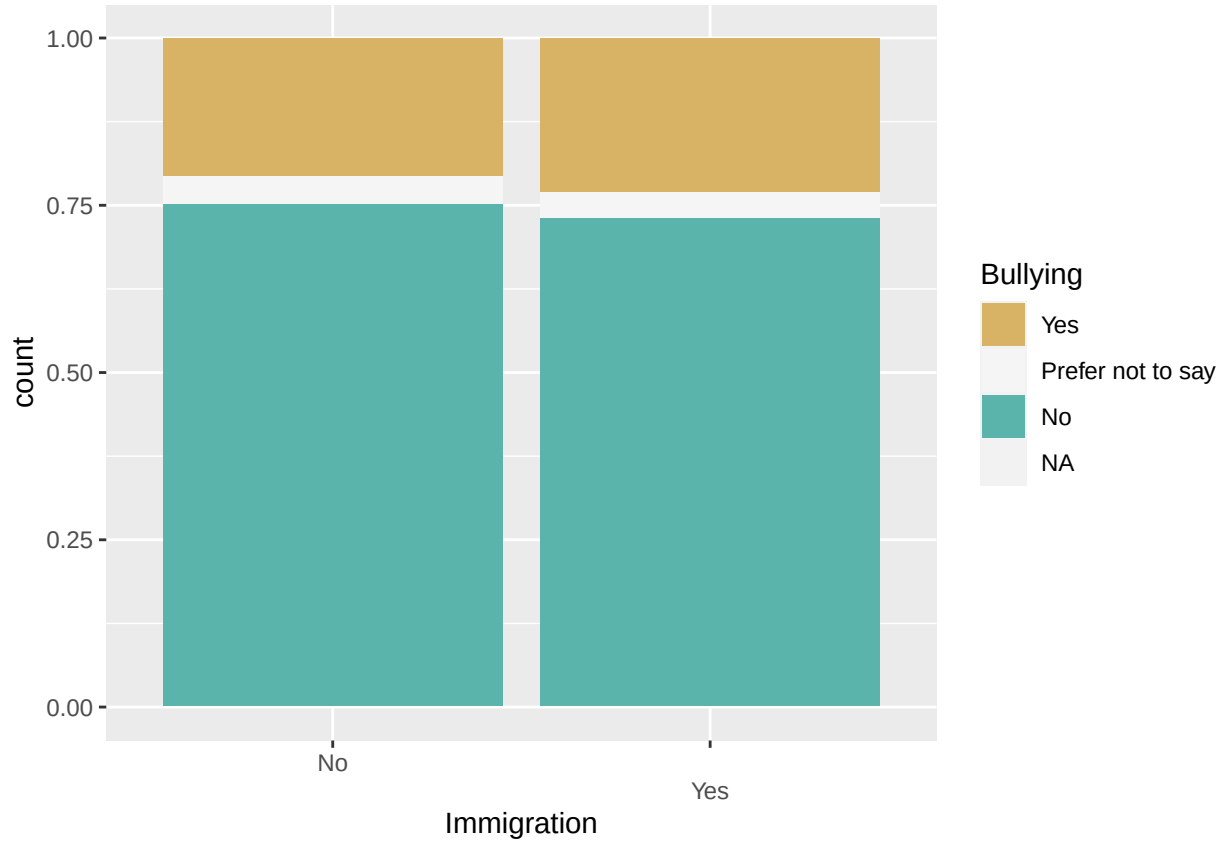
```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationBullying <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
#graphdfImmigrationFeelProgPrepMixIndAcad<- graphdfImmigrationFeelProgPrepMixIndAcad[!(graphdfImmigrati

###Refresh old graphdf1gender with data_new1 info
#graphdfImmigrationFeelProgPrepMixIndAcad <- graphdfImmigrationFeelProgPrepMixIndAcad

###Set variable order
graphdfImmigrationBullying$Bullying <- factor(graphdfImmigrationBullying$Bullying , levels=c("Yes", "Pr

##Plot
ggplot(data = graphdfImmigrationBullying) +
  geom_bar(mapping = aes(x = Immigration, fill = Bullying), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



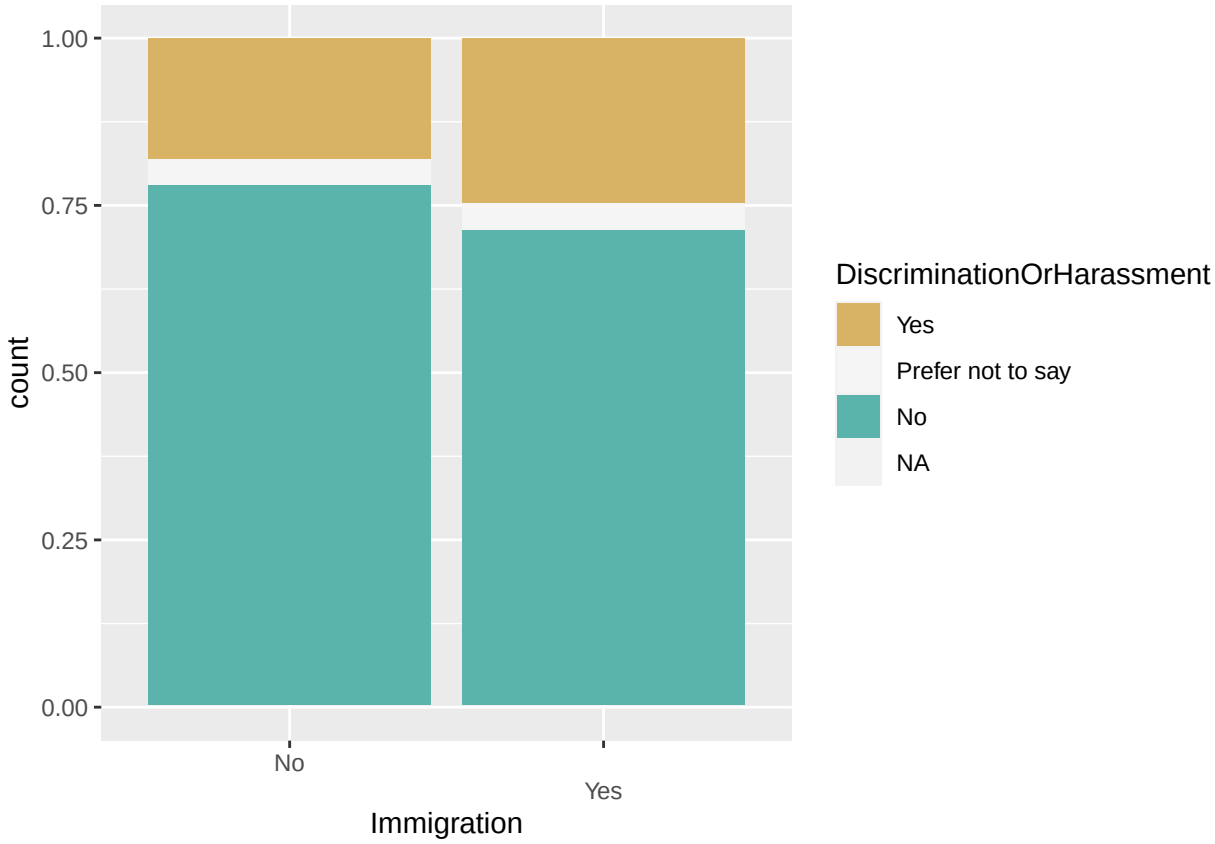
Immigration Against Experienced discrimination or harassment Column: DiscriminationOrHarassment Data frame: graphdfImmigrationDiscriminationOrHarassment

```
###New data frame, to ensure there are no issues down the line
graphdfImmigrationDiscriminationOrHarassment <- dfImm
###Keep all values that are not containing "Unsure/Not applicable"
#graphdfImmigrationDiscriminationOrHarassment<- graphdfImmigrationDiscriminationOrHarassment[!(graphdfI

###Refresh old graphdf1gender with data_new1 info
#graphdfImmigrationDiscriminationOrHarassment <- graphdfImmigrationDiscriminationOrHarassment

###Set variable order
graphdfImmigrationDiscriminationOrHarassment$DiscriminationOrHarassment <- factor(graphdfImmigrationDis
##Plot
ggplot(data = graphdfImmigrationDiscriminationOrHarassment) +
  geom_bar(mapping = aes(x = Immigration, fill = DiscriminationOrHarassment), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```

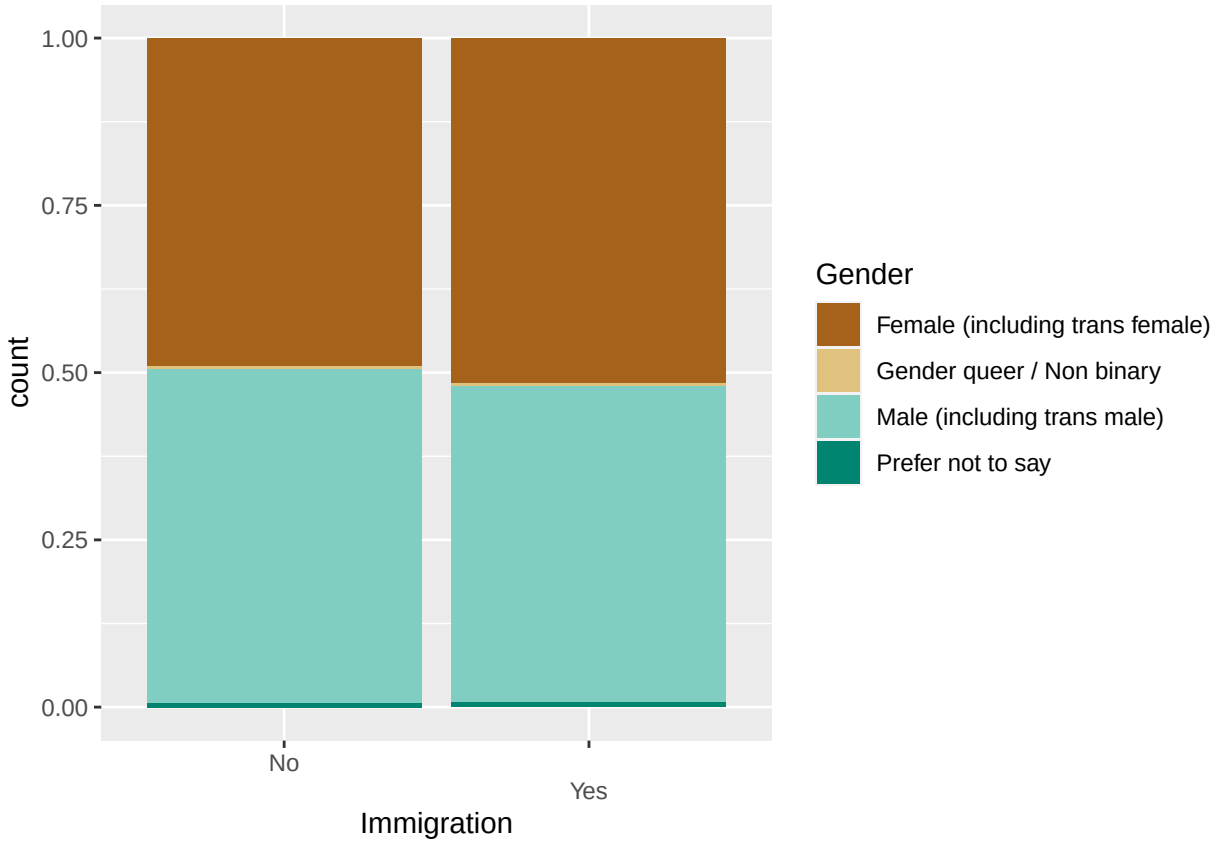


Immigration Against Gender Column: Gender Data frame: graphdfGenderAndImmigration

```
###New data frame, to ensure there are no issues down the line
graphdfGenderAndImmigration <- dfImm

##Plot
ggplot(data = graphdfGenderAndImmigration) +
  geom_bar(mapping = aes(x = Immigration, fill = Gender), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Analysis against Training and Readiness (Q50) Responses and Other Selections

Data Selection, Renaming, And Compillation

Gender count and gender proportion analysis, and data frame creation

```
###Output A Table and dataframe with the values needed
tableGender <- table(df11['Gender'])
genderCount<- as.data.frame(tableGender)

###Proportional Analysis
genderProportion <- as.data.frame(table(df11$Gender)/length(df11$Gender))
###Display Output genderProportion
genderProportion
```

```
##                               Var1        Freq
## 1 Female (including trans female) 0.499706400
## 2       Gender queer / Non binary 0.004697592
## 3     Male (including trans male) 0.489136817
## 4                Prefer not to say 0.006459190
```

Creation of different data frames for analysis by gender, against training efficacy for certain tasks(Q50). Not used here but will be used later on.

```
df12 <- df11[c(201,202,203,204,205,206,207,208,209,210,211,212,213,214,215,216,242)]

df12M <- df12[df12$'Gender' == 'Male (including trans male)',]
df12F <- df12[df12$'Gender' == 'Female (including trans female)',]
df12GQNB <- df12[df12$'Gender' == 'Gender queer / Non binary',]
```

A workflow for analysis of gender by creating seperate dataframes, using data frame with only "Male (including trans male)" as an example.

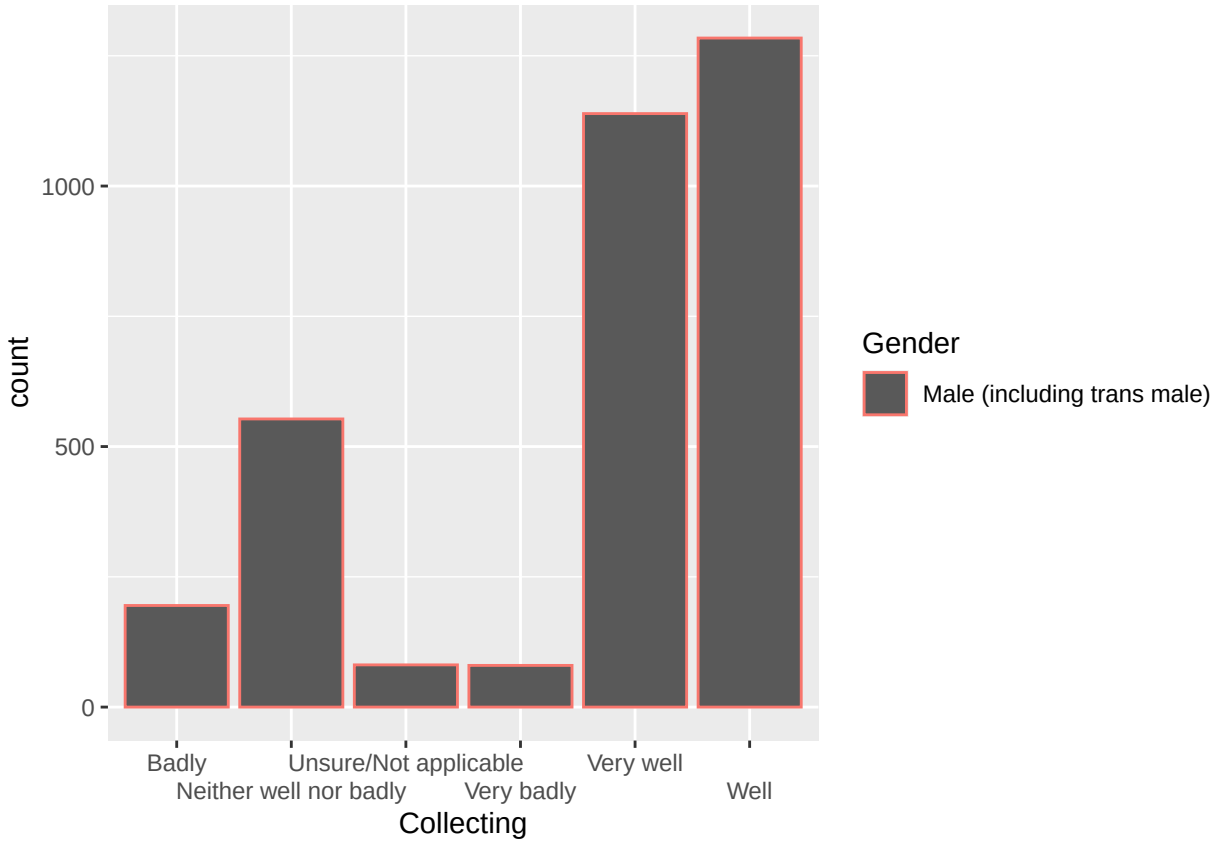
```
###Select relevant columns/variables

### Rename Variable/Column to be analyzed, the dependent, by index
df13M <- df12M
df14M <- df13M
df14M <- names(df13M)[1] <- "Collecting"
df14M <- df13M


###Plot the bit
#ggplot(df14M, aes( Collecting, colour = Gender)) +   geom_bar() ###Just another way to plot the data

barplotMCollectingData <- ggplot(df14M, aes( Collecting, colour = Gender)) +
  geom_bar()+
  scale_x_discrete(guide = guide_axis(n.dodge=2))

###Display Plot
barplotMCollectingData
```



New data frame for gender and transformation of previous dataframe into one that will plug in directly. I wrote the code for gender first, but dfImm allows for a larger number of analysies, so I will just copy it in, and it still ought to work.

```
dfGen <- dfImm

df15 <- dfGen
df <- df15
```

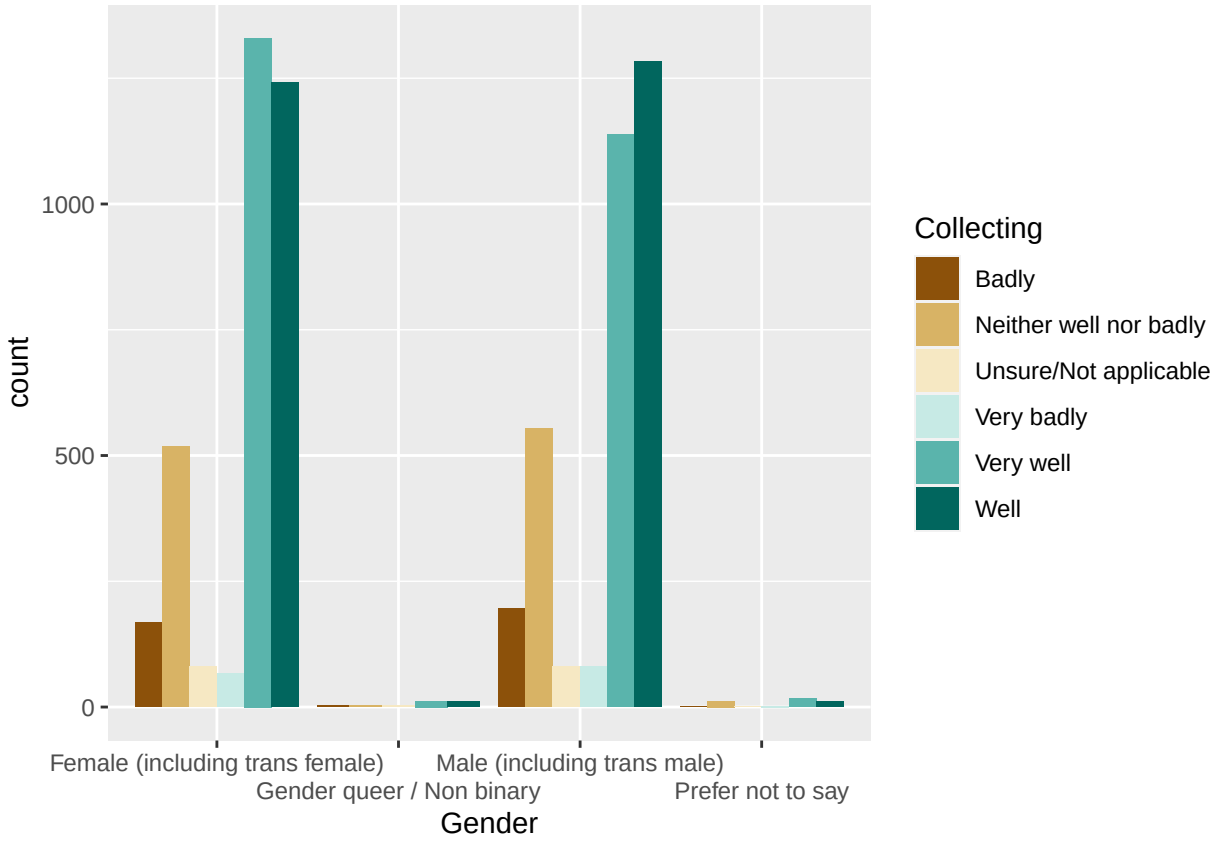
Graphs for Gender

####Gender against Collecting Data

first graph for gender against collecting data, not incredibly interesting, and also hampered by small sample size of Gender Queer/Non Binary and Prefer Not To Say

```
groupedBarGenderCollecting <- ggplot(data = df) +
  geom_bar(mapping = aes(x = Gender, fill = Collecting), position = "dodge")+
  scale_colour_brewer(palette = "BrBG") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))


groupedBarGenderCollecting
```



A better proportional graph. This chunk also contains the code to reorder variables manually.

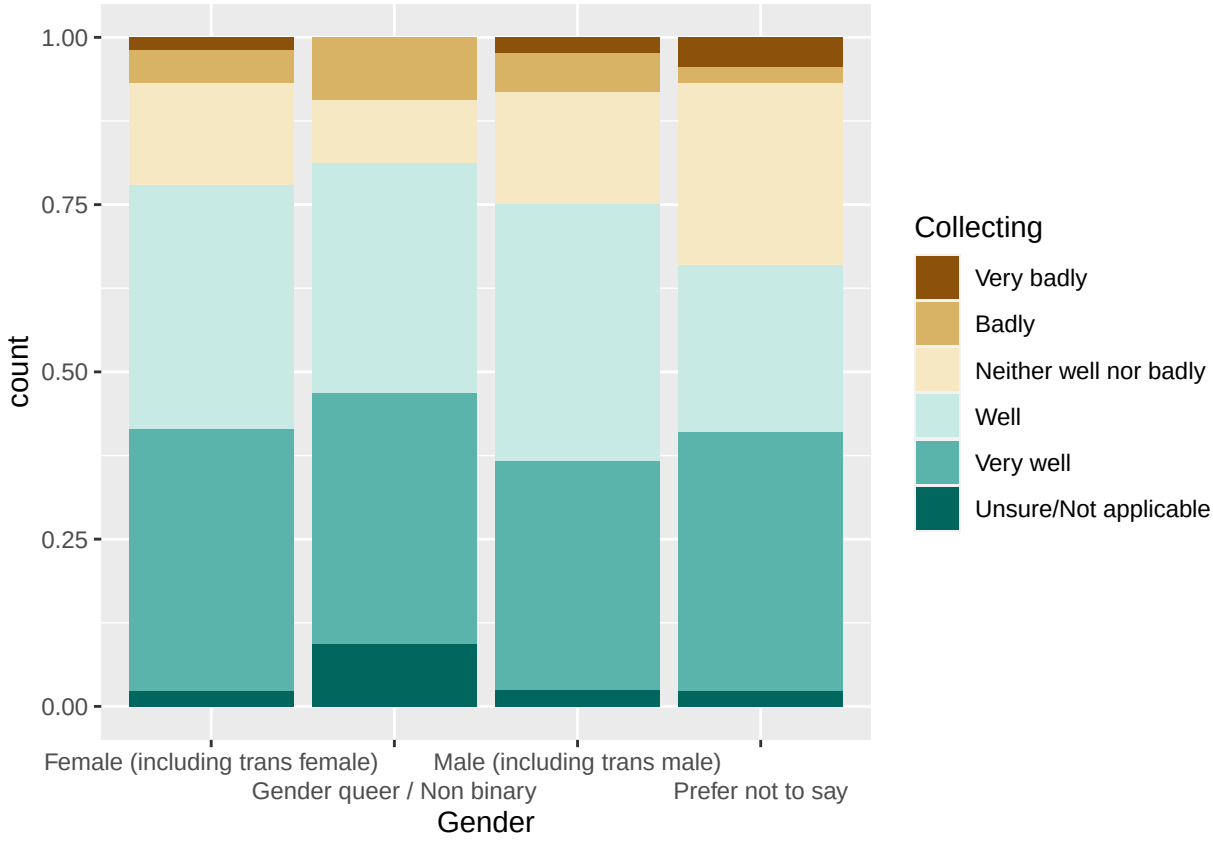
```
data_new <- df
df <- data_new

df$Collecting <- factor(df$Collecting ,
                        levels=c("Very badly", "Badly", "Neither well nor badly",
                                 "Well", "Very well", "Unsure/Not applicable"))

data_new <- df
df <- data_new

ggplot(data = data_new) +
  geom_bar(mapping = aes(x = Gender, fill = Collecting), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Not bad, however the not sure/not applicable bit is distorting our visual analysis.

Same proportional graph, without the "Unsure/Not applicable" value

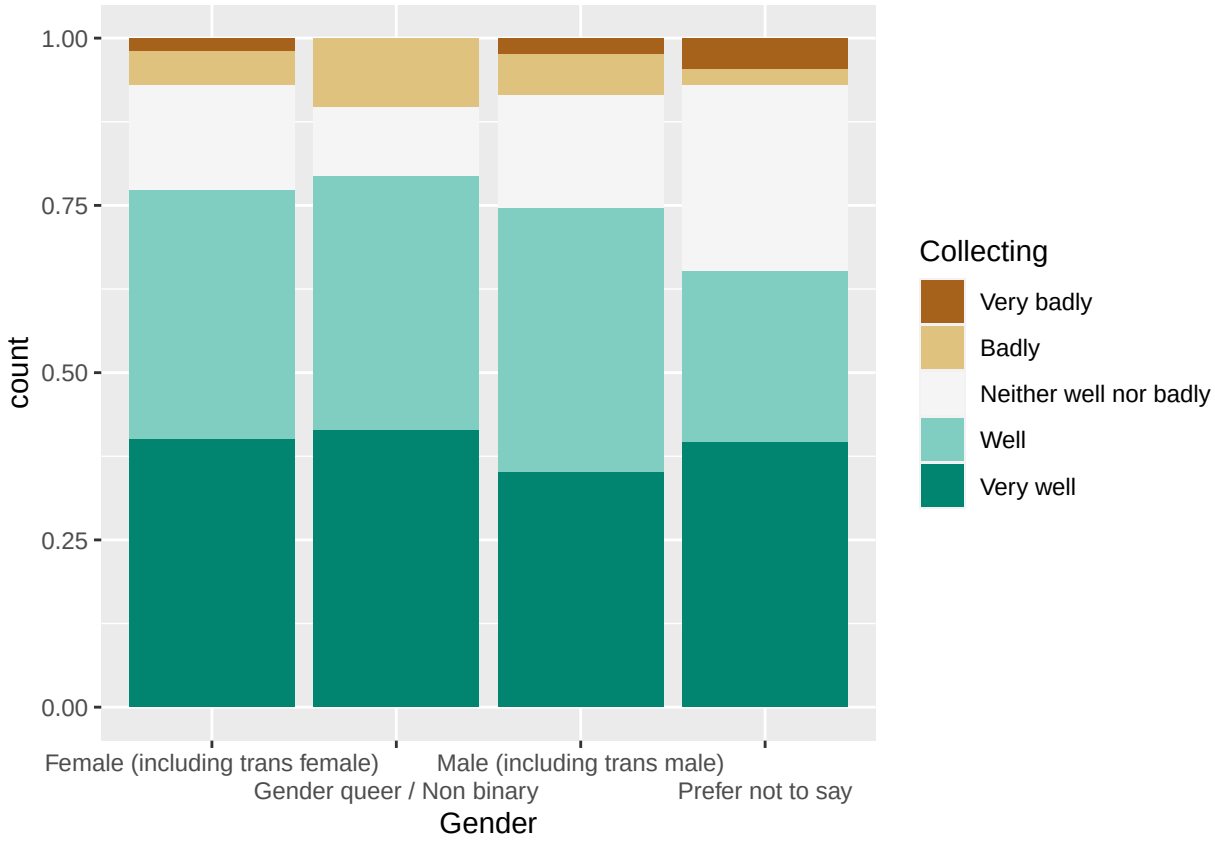
```
###New data frame, to ensure there are no issues down the line
graphdf1gender <- df
###Keep all values that are not containing "Unsure/Not applicable"
data_new1 <- graphdf1gender[!(graphdf1gender$Collecting == "Unsure/Not applicable"),]

###Refresh old graphdf1gender with data_new1 info
graphdf1gender <- data_new1

###Set variable order
graphdf1gender$Collecting <- factor(graphdf1gender$Collecting , levels=c("Very badly", "Badly", "Neithe

##Plot
ggplot(data = data_new1) +
  geom_bar(mapping = aes(x = Gender, fill = Collecting), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Analyzing Data Column: Analyzing Data Frame: graphdf1genderAnalyzing

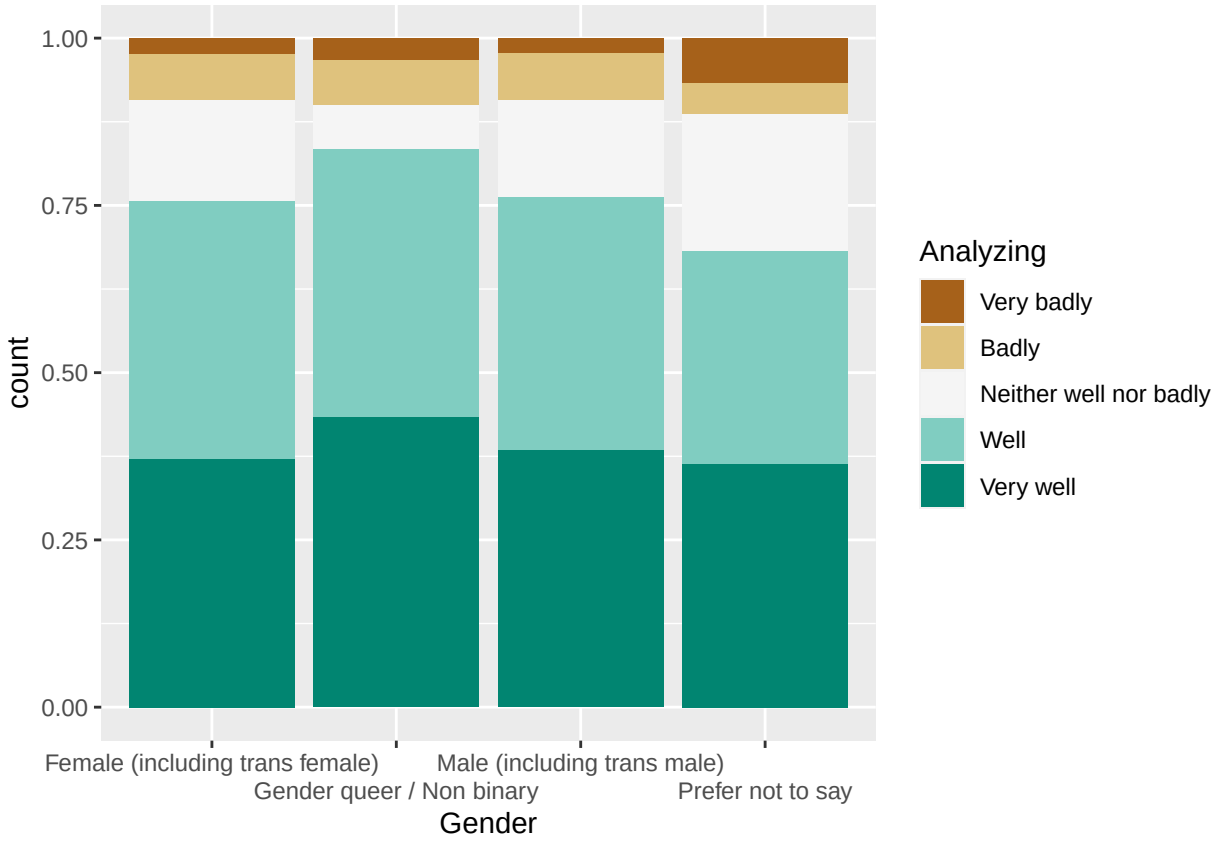
```
###New data frame, to ensure there are no issues down the line
graphdf1genderAnalyzing <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderAnalyzing<- graphdf1genderAnalyzing[!(graphdf1genderAnalyzing$Analyzing == "Unsure/Not ap

###Refresh old graphdf1gender with data_new1 info
graphdf1genderAnalyzing <- graphdf1genderAnalyzing

###Set variable order
graphdf1genderAnalyzing$Analyzing <- factor(graphdf1genderAnalyzing$Analyzing , levels=c("Very badly", "

##Plot
ggplot(data = graphdf1genderAnalyzing) +
  geom_bar(mapping = aes(x = Gender, fill = Analyzing), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Designing robust reproducible experiments Column: Designing Data Frame: graphdf1genderDesigning

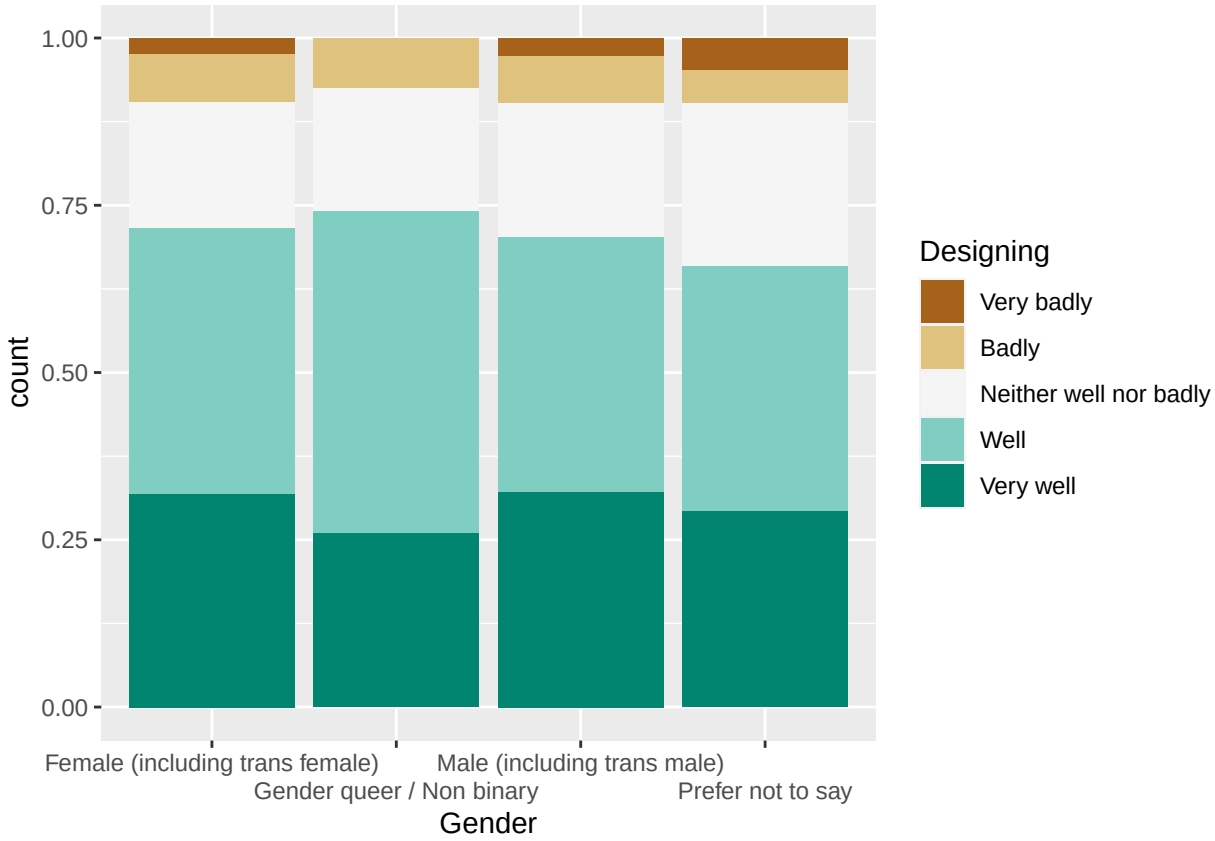
```
###New data frame, to ensure there are no issues down the line
graphdf1genderDesigning <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderDesigning<- graphdf1genderDesigning[!(graphdf1genderDesigning$Designing == "Unsure/Not ap

###Refresh old graphdf1gender with data_new1 info
graphdf1genderDesigning <- graphdf1genderDesigning

###Set variable order
graphdf1genderDesigning$Designing <- factor(graphdf1genderDesigning$Designing , levels=c("Very badly",

##Plot
ggplot(data = graphdf1genderDesigning) +
  geom_bar(mapping = aes(x = Gender, fill = Designing), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Writing a paper for publication in a peer-reviewed journal Column: Writing
Data Frame: graphdf1genderWriting

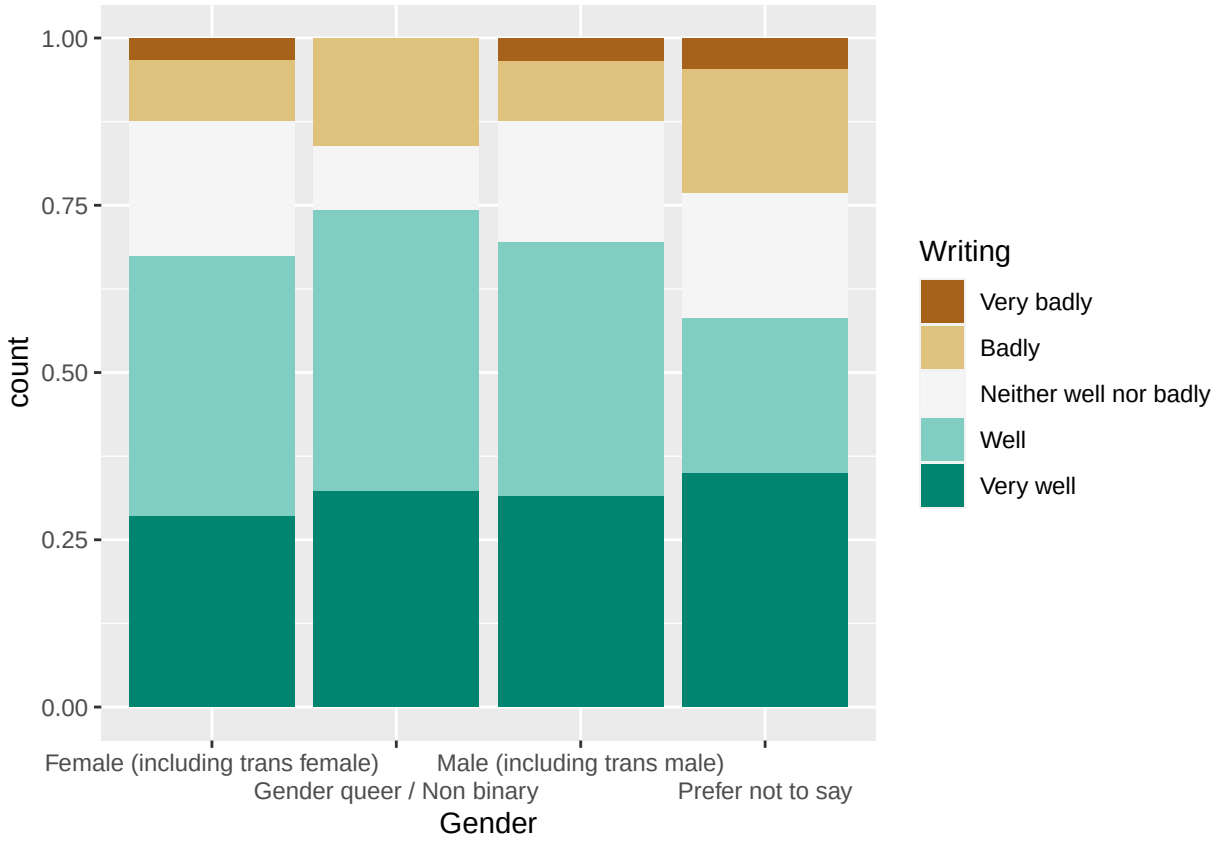
```
###New data frame, to ensure there are no issues down the line
graphdf1genderWriting <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderWriting<- graphdf1genderWriting[!(graphdf1genderWriting$Writing == "Unsure/Not applicable

###Refresh old graphdf1gender with data_new1 info
graphdf1genderWriting <- graphdf1genderWriting

###Set variable order
graphdf1genderWriting$Writing <- factor(graphdf1genderWriting$Writing , levels=c("Very badly", "Badly",

##Plot
ggplot(data = graphdf1genderWriting) +
  geom_bar(mapping = aes(x = Gender, fill = Writing), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Developing resilience to manage rejection by a peer review panel Column: DevResistance Data Frame: graphdf1genderDevResistance

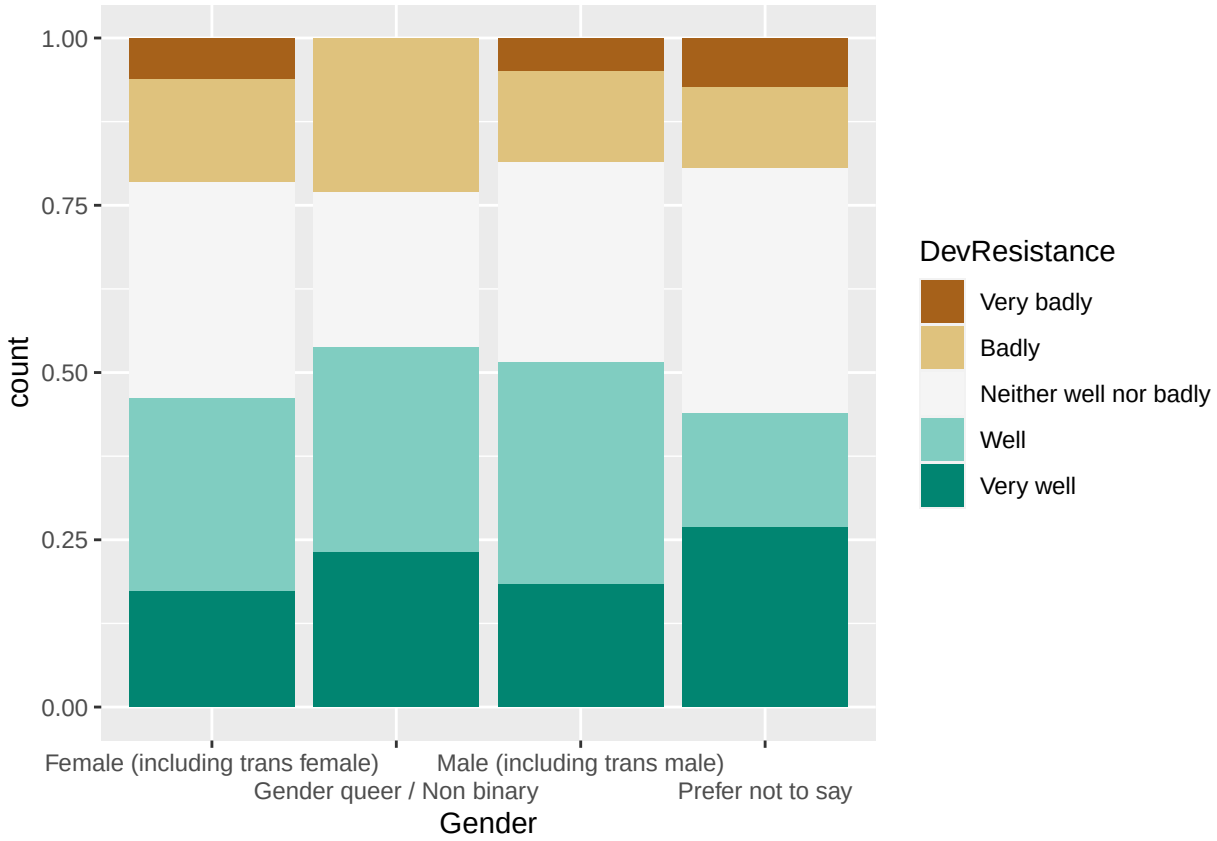
```
###New data frame, to ensure there are no issues down the line
graphdf1genderDevResistance <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderDevResistance<- graphdf1genderDevResistance[!(graphdf1genderDevResistance$DevResistance ==

###Refresh old graphdf1gender with data_new1 info
graphdf1genderDevResistance <- graphdf1genderDevResistance

###Set variable order
graphdf1genderDevResistance$DevResistance <- factor(graphdf1genderDevResistance$DevResistance , levels=

##Plot
ggplot(data = graphdf1genderDevResistance) +
  geom_bar(mapping = aes(x = Gender, fill = DevResistance), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Presenting findings to a specialist audience Column: PresentingSpecialist Data Frame: graphdf1genderPresentingSpecialist

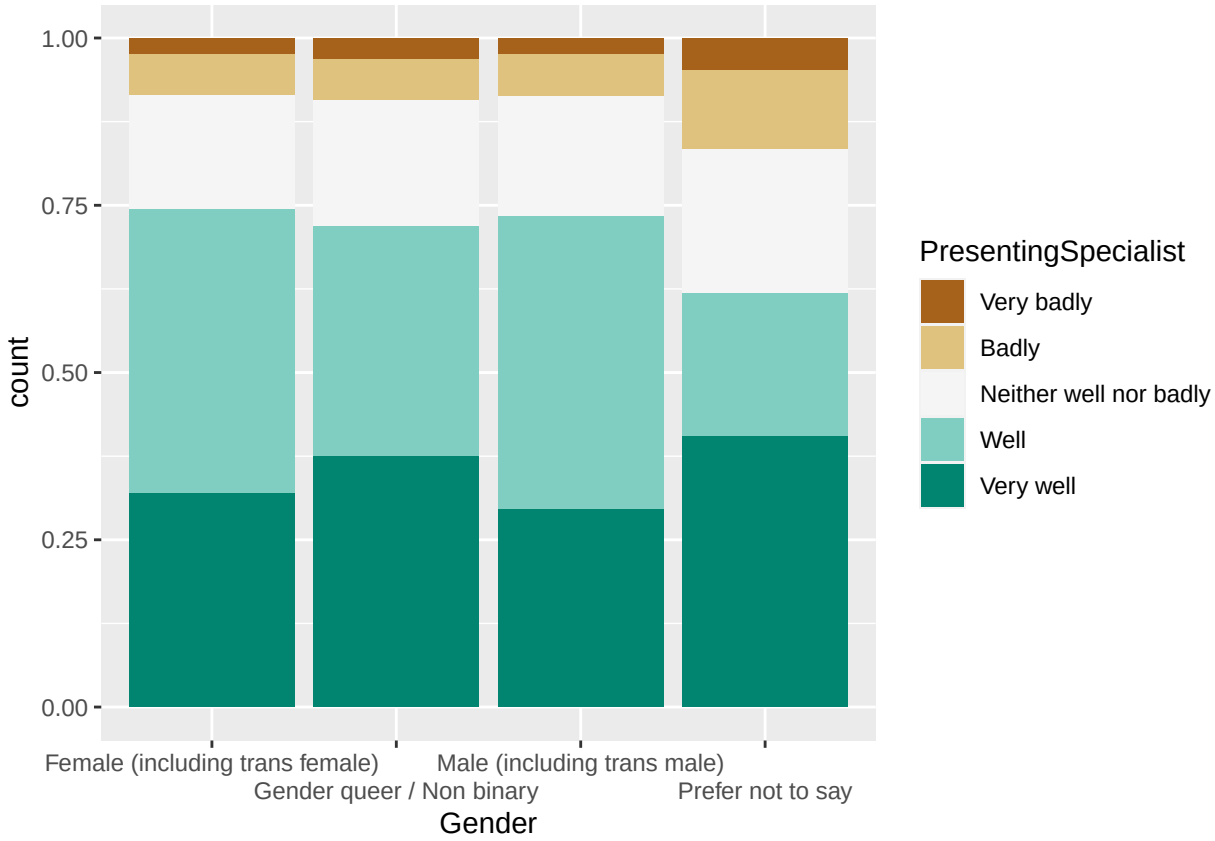
```
###New data frame, to ensure there are no issues down the line
graphdf1genderPresentingSpecialist <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderPresentingSpecialist<- graphdf1genderPresentingSpecialist[!(graphdf1genderPresentingSpeci

###Refresh old graphdf1gender with data_new1 info
graphdf1genderPresentingSpecialist <- graphdf1genderPresentingSpecialist

###Set variable order
graphdf1genderPresentingSpecialist$PresentingSpecialist <- factor(graphdf1genderPresentingSpecialist$Pr

##Plot
ggplot(data = graphdf1genderPresentingSpecialist) +
  geom_bar(mapping = aes(x = Gender, fill = PresentingSpecialist), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Presenting findings to a non-specialist (public) audience Column: Presenting-Public Data Frame: graphdf1genderPresentingPublic

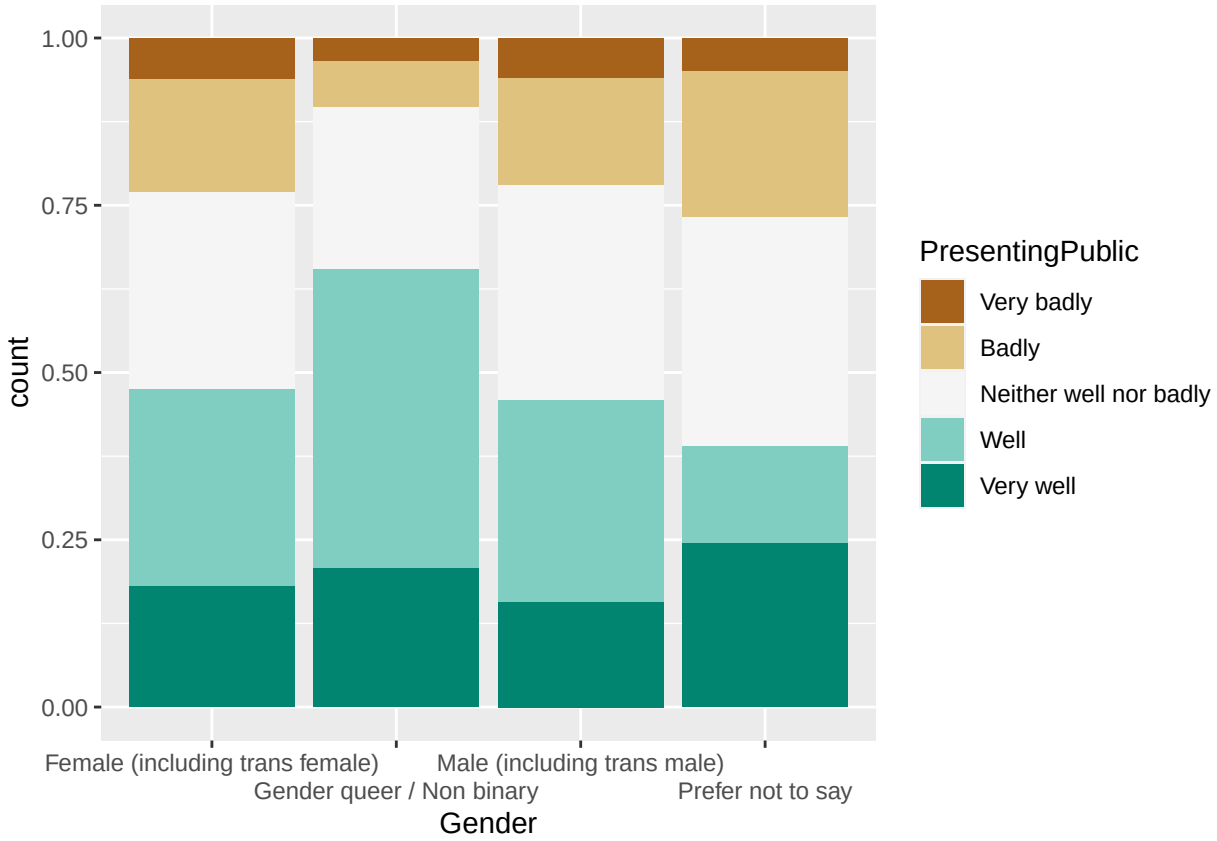
```
###New data frame, to ensure there are no issues down the line
graphdf1genderPresentingPublic <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderPresentingPublic<- graphdf1genderPresentingPublic[!(graphdf1genderPresentingPublic$Presen

###Refresh old graphdf1gender with data_new1 info
graphdf1genderPresentingPublic <- graphdf1genderPresentingPublic

###Set variable order
graphdf1genderPresentingPublic$PresentingPublic <- factor(graphdf1genderPresentingPublic$PresentingPubli

##Plot
ggplot(data = graphdf1genderPresentingPublic) +
  geom_bar(mapping = aes(x = Gender, fill = PresentingPublic), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Finding a Applying for funding Column: ApplyingFunding Data Frame: graphdf1genderApplyingFunding

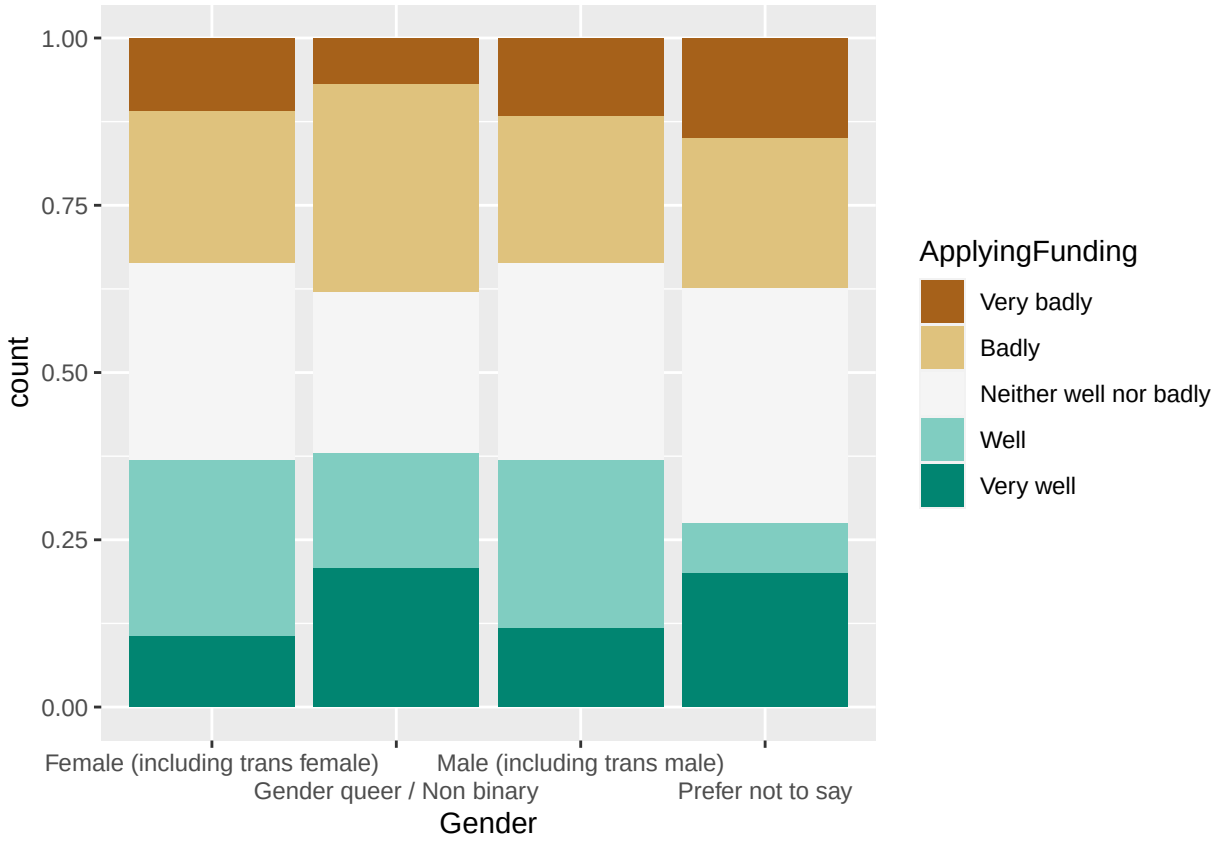
```
###New data frame, to ensure there are no issues down the line
graphdf1genderApplyingFunding <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderApplyingFunding<- graphdf1genderApplyingFunding[!(graphdf1genderApplyingFunding$ApplyingF

###Refresh old graphdf1gender with data_new1 info
graphdf1genderApplyingFunding <- graphdf1genderApplyingFunding

###Set variable order
graphdf1genderApplyingFunding$ApplyingFunding <- factor(graphdf1genderApplyingFunding$ApplyingFunding ,

##Plot
ggplot(data = graphdf1genderApplyingFunding) +
  geom_bar(mapping = aes(x = Gender, fill = ApplyingFunding), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Finding a satisfying career Column: SatisCareer Data Frame: graphdf1genderSatisCareer

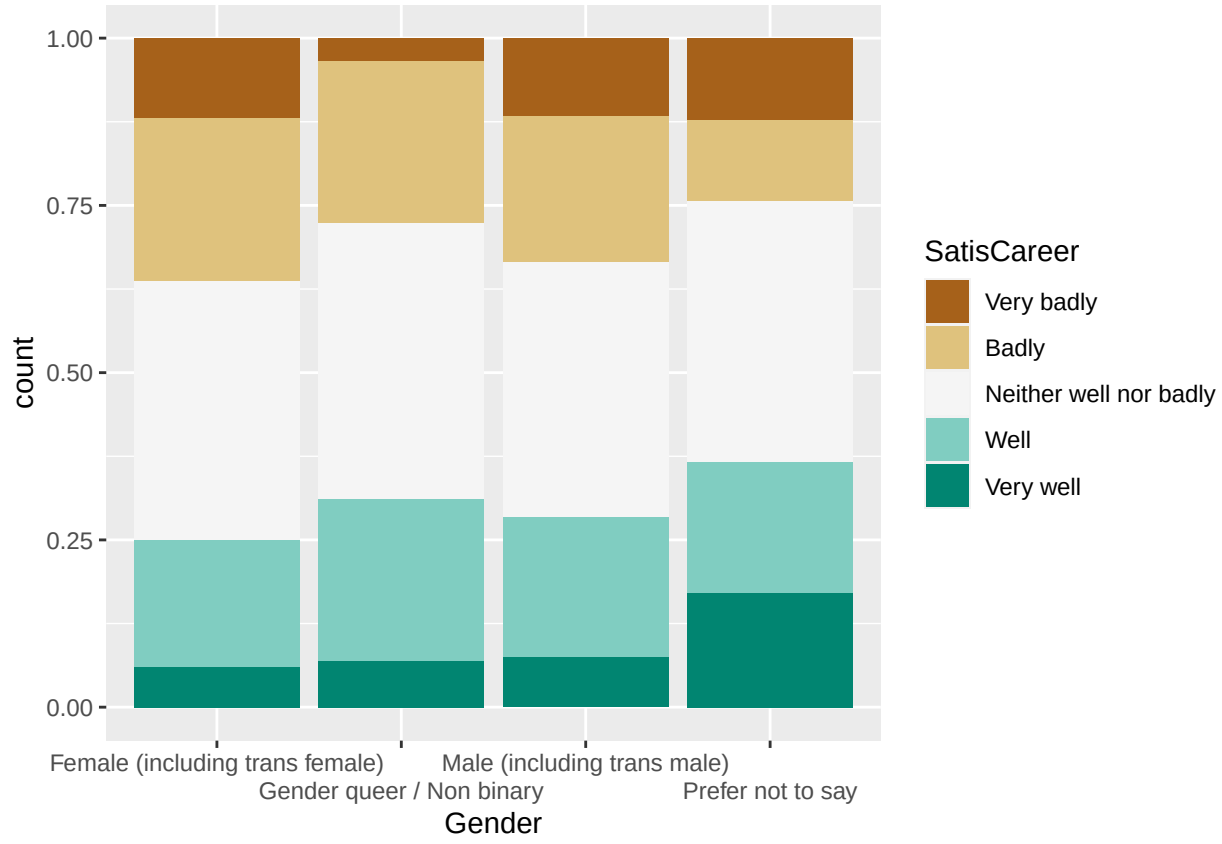
```
###New data frame, to ensure there are no issues down the line
graphdf1genderSatisCareer <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderSatisCareer<- graphdf1genderSatisCareer[!(graphdf1genderSatisCareer$SatisCareer == "Unsur

###Refresh old graphdf1gender with data_new1 info
graphdf1genderSatisCareer <- graphdf1genderSatisCareer

###Set variable order
graphdf1genderSatisCareer$SatisCareer <- factor(graphdf1genderSatisCareer$SatisCareer , levels=c("Very

##Plot
ggplot(data = graphdf1genderSatisCareer) +
  geom_bar(mapping = aes(x = Gender, fill = SatisCareer), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Managing complex projects Column: MngCompProj Data Frame: graphdf1genderMngCompProj

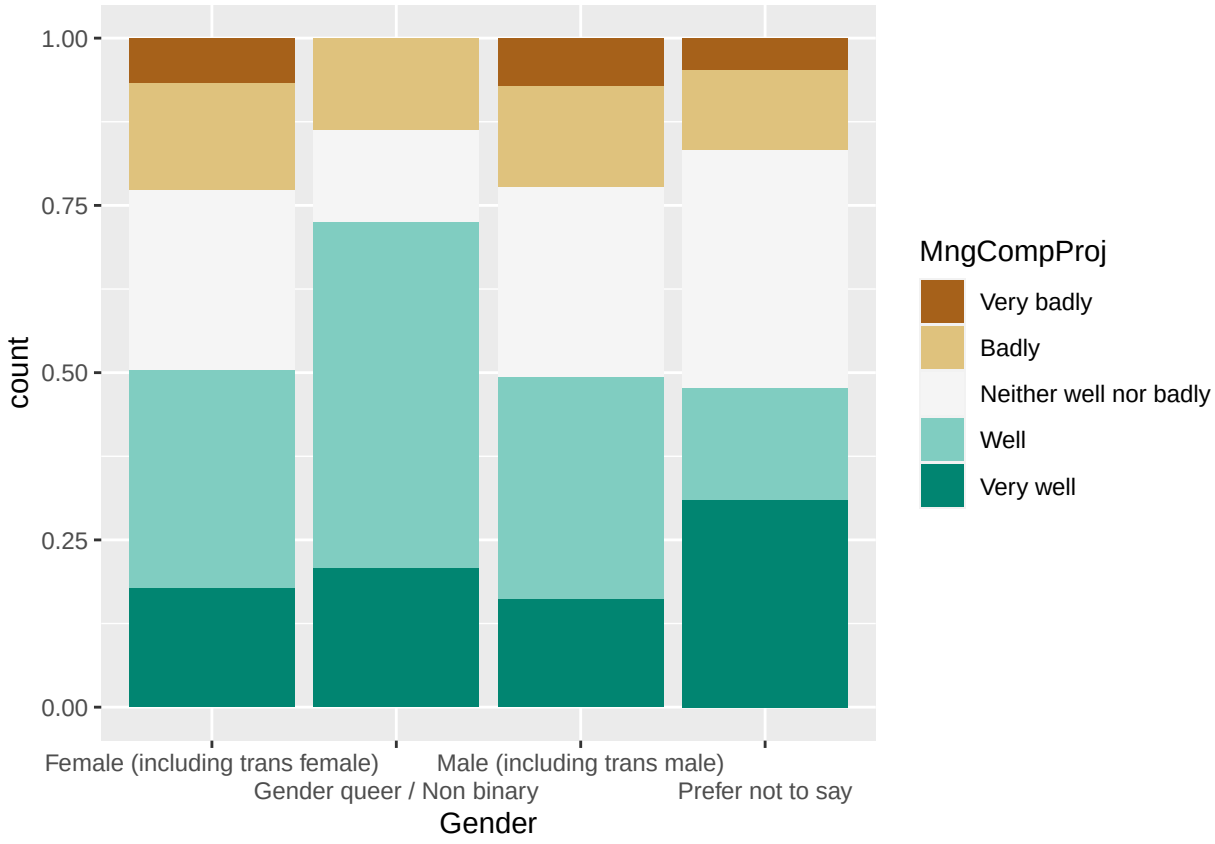
```
###New data frame, to ensure there are no issues down the line
graphdf1genderMngCompProj <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderMngCompProj<- graphdf1genderMngCompProj[!(graphdf1genderMngCompProj$MngCompProj == "Unsur

###Refresh old graphdf1gender with data_new1 info
graphdf1genderMngCompProj <- graphdf1genderMngCompProj

###Set variable order
graphdf1genderMngCompProj$MngCompProj <- factor(graphdf1genderMngCompProj$MngCompProj , levels=c("Very

##Plot
ggplot(data = graphdf1genderMngCompProj) +
  geom_bar(mapping = aes(x = Gender, fill = MngCompProj), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Developing a business plan DevBusinessPlan Data Frame: graphdf1genderDevBusinessPlan

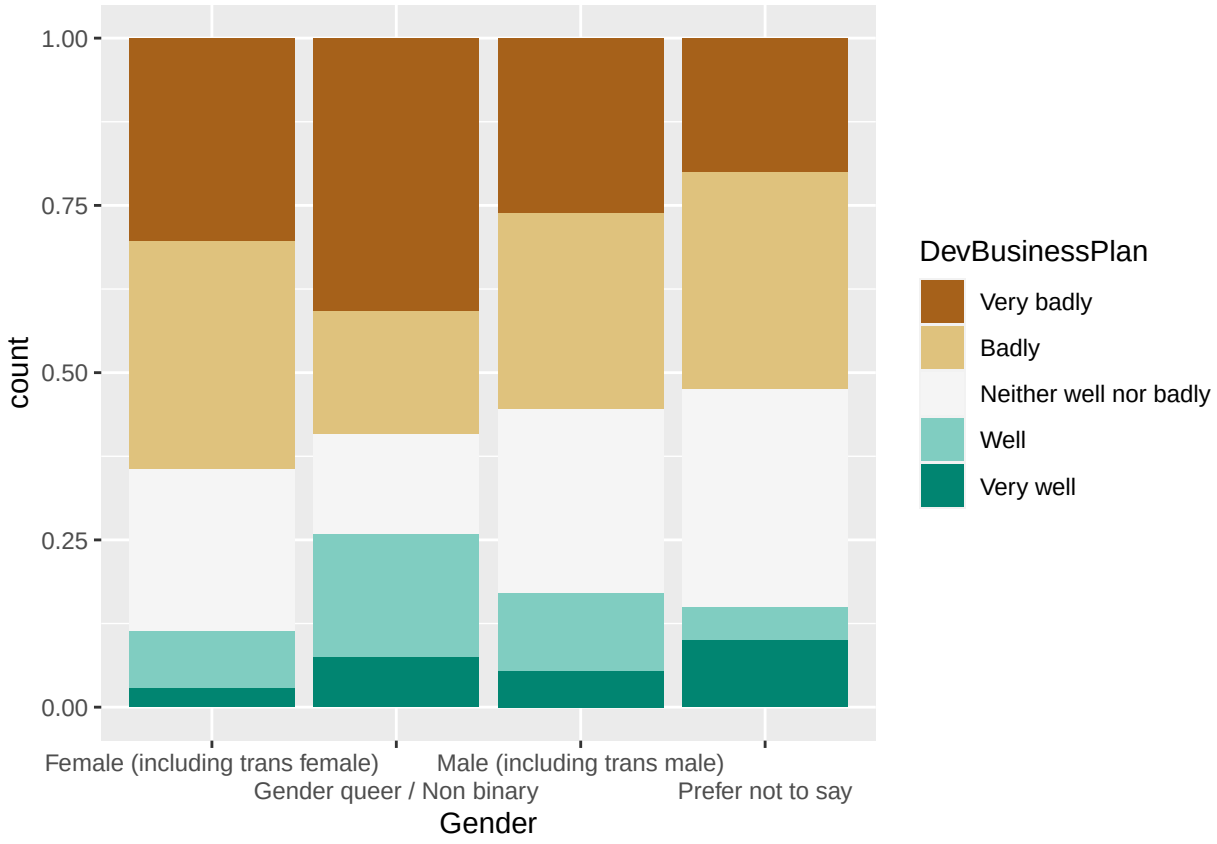
```
###New data frame, to ensure there are no issues down the line
graphdf1genderDevBusinessPlan <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderDevBusinessPlan<- graphdf1genderDevBusinessPlan[!(graphdf1genderDevBusinessPlan$DevBusines

###Refresh old graphdf1gender with data_new1 info
graphdf1genderDevBusinessPlan <- graphdf1genderDevBusinessPlan

###Set variable order
graphdf1genderDevBusinessPlan$DevBusinessPlan <- factor(graphdf1genderDevBusinessPlan$DevBusinessPlan ,

##Plot
ggplot(data = graphdf1genderDevBusinessPlan) +
  geom_bar(mapping = aes(x = Gender, fill = DevBusinessPlan), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Managing people Column: MngPeople Data Frame: graphdf1genderMngPeople

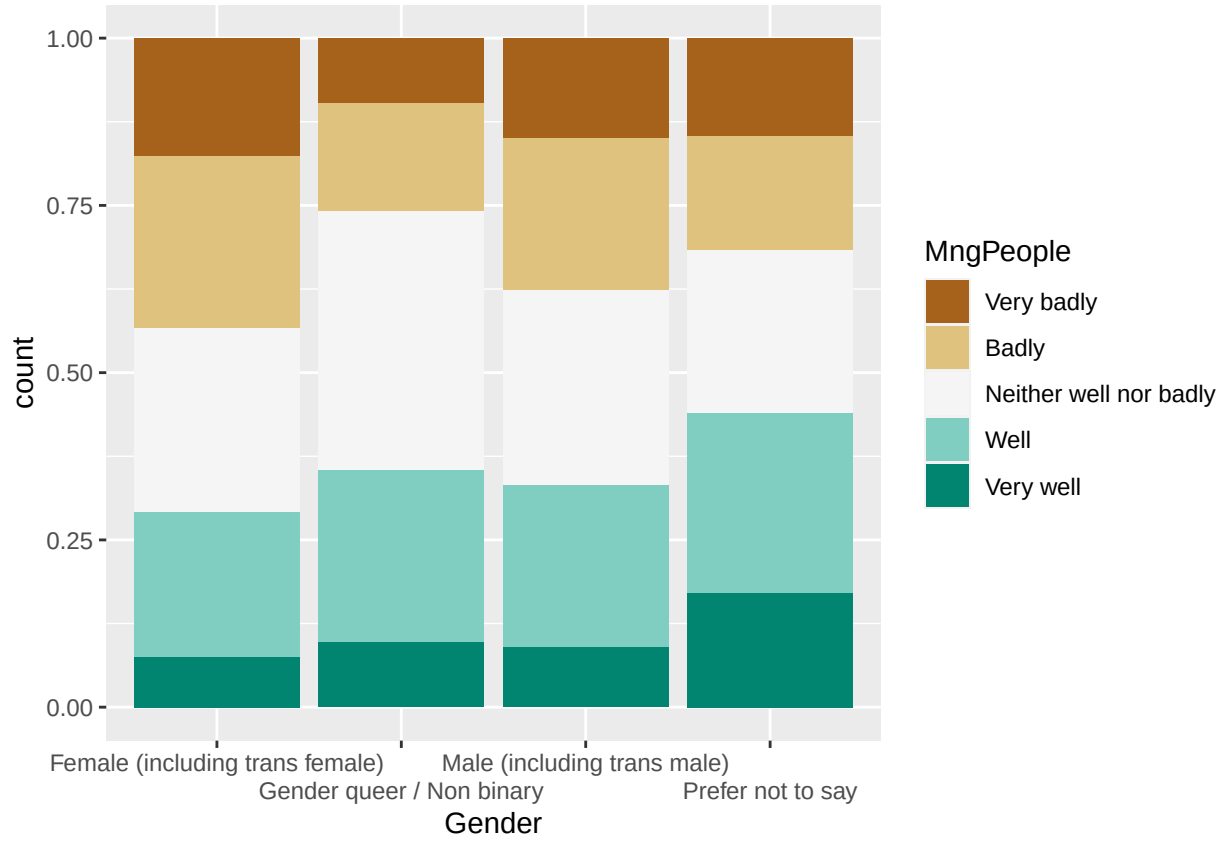
```
###New data frame, to ensure there are no issues down the line
graphdf1genderMngPeople <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderMngPeople<- graphdf1genderMngPeople[!(graphdf1genderMngPeople$MngPeople == "Unsure/Not ap

###Refresh old graphdf1gender with data_new1 info
graphdf1genderMngPeople <- graphdf1genderMngPeople

###Set variable order
graphdf1genderMngPeople$MngPeople <- factor(graphdf1genderMngPeople$MngPeople , levels=c("Very badly",

##Plot
ggplot(data = graphdf1genderMngPeople) +
  geom_bar(mapping = aes(x = Gender, fill = MngPeople), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against Managing a large operational budget Column: MngLargeBudget Data Frame: graphdf1genderMngLargeBudget

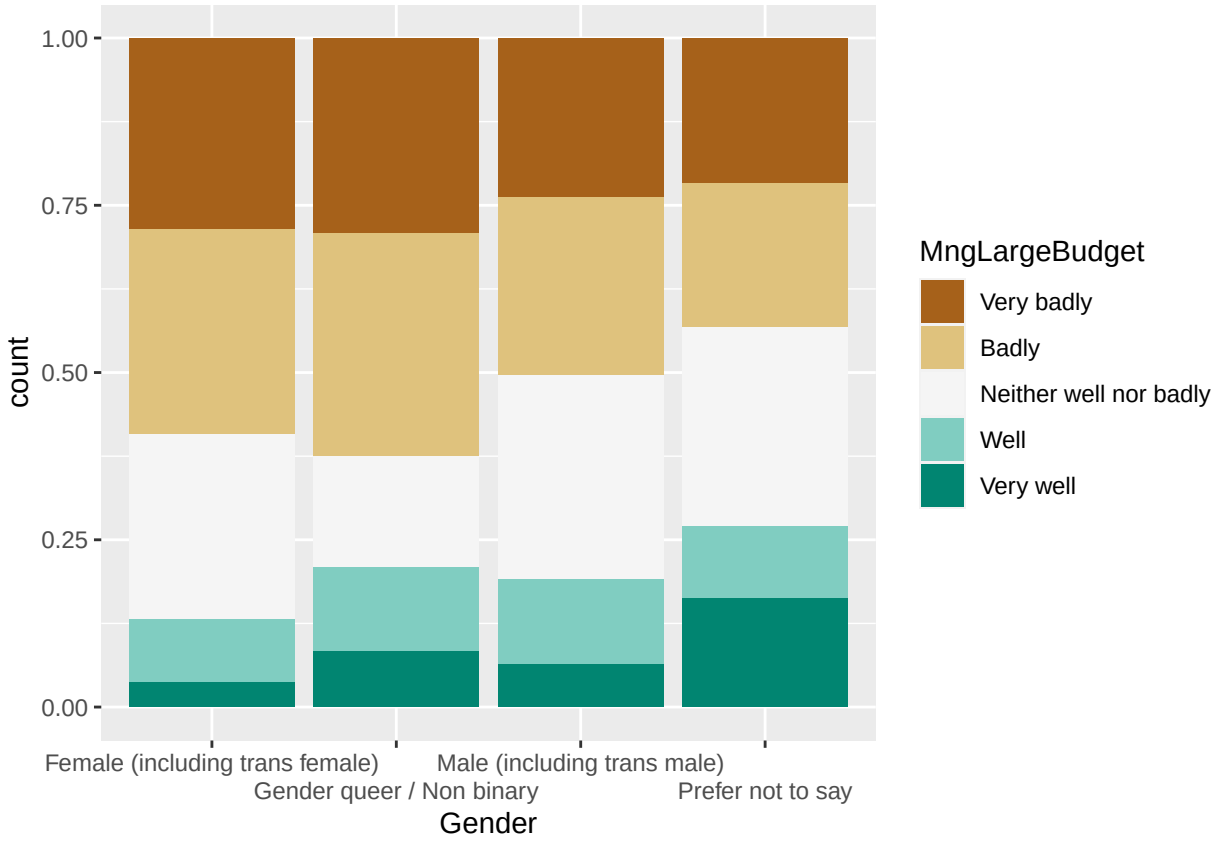
```
###New data frame, to ensure there are no issues down the line
graphdf1genderMngLargeBudget <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderMngLargeBudget<- graphdf1genderMngLargeBudget[!(graphdf1genderMngLargeBudget$MngLargeBudg

###Refresh old graphdf1gender with data_new1 info
graphdf1genderMngLargeBudget <- graphdf1genderMngLargeBudget

###Set variable order
graphdf1genderMngLargeBudget$MngLargeBudget <- factor(graphdf1genderMngLargeBudget$MngLargeBudget , lev

##Plot
ggplot(data = graphdf1genderMngLargeBudget) +
  geom_bar(mapping = aes(x = Gender, fill = MngLargeBudget), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against I feel that my programme is preparing me well for a research career Column: FeelProgPrepResearch Data Frame: graphdf1genderFeelProgPrepResearch

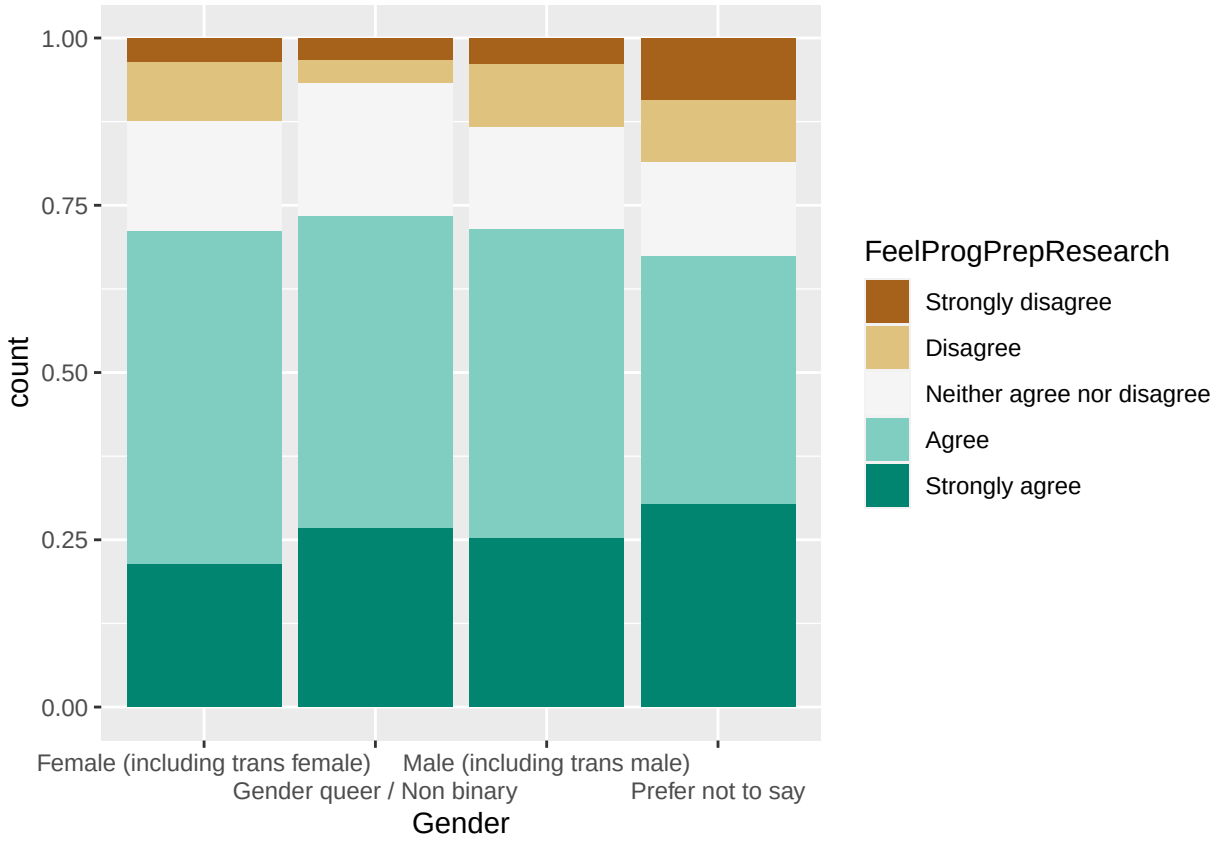
```
###New data frame, to ensure there are no issues down the line
graphdf1genderFeelProgPrepResearch <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderFeelProgPrepResearch<- graphdf1genderFeelProgPrepResearch[!(graphdf1genderFeelProgPrepRes

###Refresh old graphdf1gender with data_new1 info
graphdf1genderFeelProgPrepResearch <- graphdf1genderFeelProgPrepResearch

###Set variable order
graphdf1genderFeelProgPrepResearch$FeelProgPrepResearch <- factor(graphdf1genderFeelProgPrepResearch$Fe

##Plot
ggplot(data = graphdf1genderFeelProgPrepResearch) +
  geom_bar(mapping = aes(x = Gender, fill = FeelProgPrepResearch), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against I feel that my programme is perparing me well for a non-research science-related career Column: FeelProgPrepScience Data Frame: graphdf1genderFeelProgPrepScience

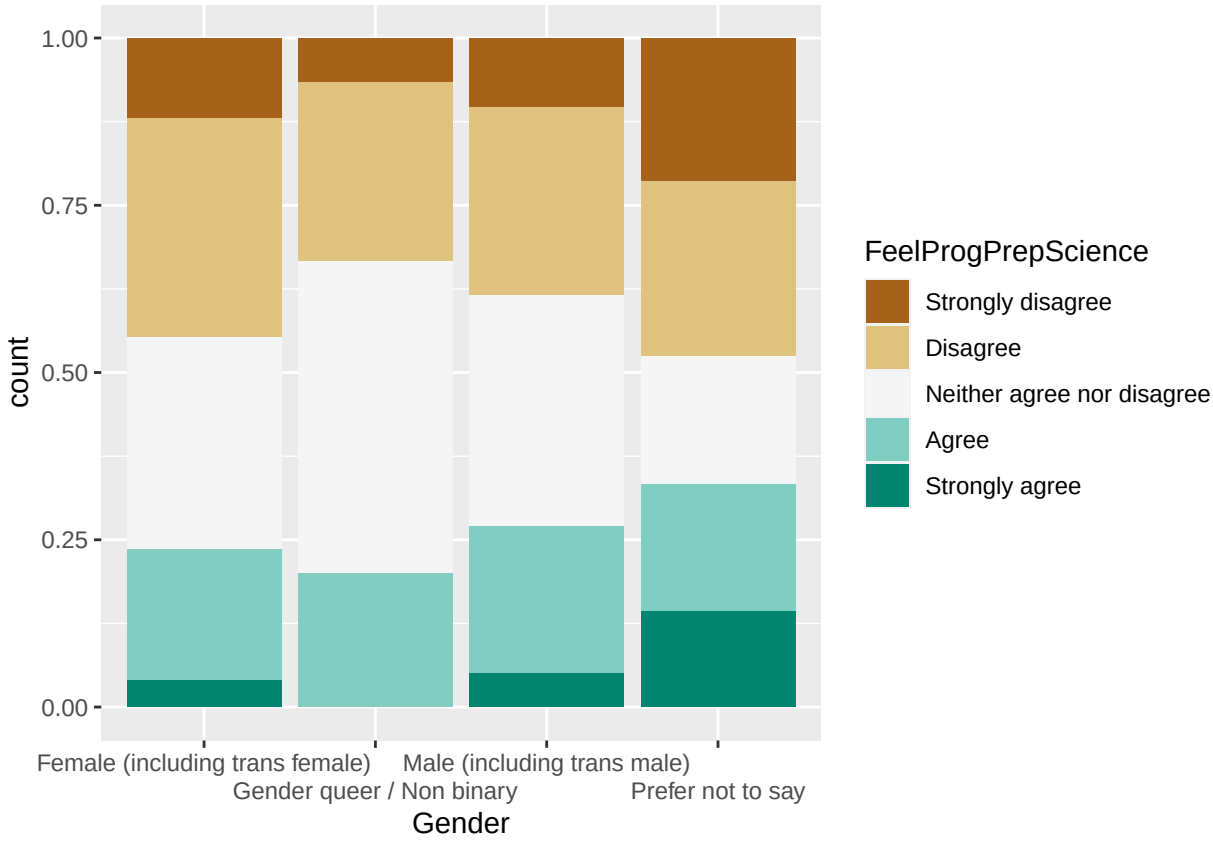
```
###New data frame, to ensure there are no issues down the line
graphdf1genderFeelProgPrepScience <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderFeelProgPrepScience<- graphdf1genderFeelProgPrepScience[!(graphdf1genderFeelProgPrepScien

###Refresh old graphdf1gender with data_new1 info
graphdf1genderFeelProgPrepScience <- graphdf1genderFeelProgPrepScience

###Set variable order
graphdf1genderFeelProgPrepScience$FeelProgPrepScience <- factor(graphdf1genderFeelProgPrepScience$FeelP

##Plot
ggplot(data = graphdf1genderFeelProgPrepScience) +
  geom_bar(mapping = aes(x = Gender, fill = FeelProgPrepScience), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Gender Against I feel that my programme is preparing me well for a career that straddles both industry and academia Column: FeelProgPrepMixIndAcad Data Frame: graphdf1genderFeelProgPrepMixIndAcad

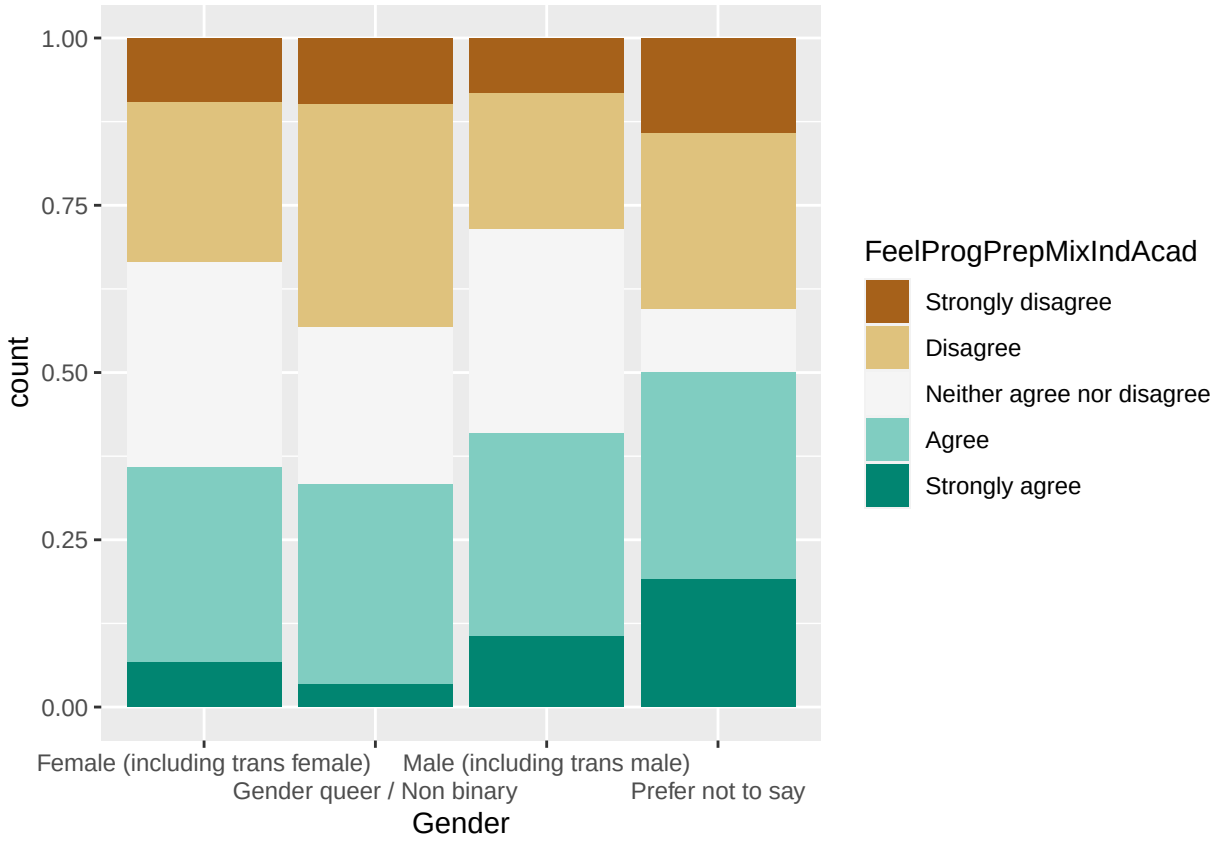
```
###New data frame, to ensure there are no issues down the line
graphdf1genderFeelProgPrepMixIndAcad <- df
###Keep all values that are not containing "Unsure/Not applicable"
graphdf1genderFeelProgPrepMixIndAcad<- graphdf1genderFeelProgPrepMixIndAcad[!(graphdf1genderFeelProgPrep

###Refresh old graphdf1gender with data_new1 info
graphdf1genderFeelProgPrepMixIndAcad <- graphdf1genderFeelProgPrepMixIndAcad

###Set variable order
graphdf1genderFeelProgPrepMixIndAcad$FeelProgPrepMixIndAcad <- factor(graphdf1genderFeelProgPrepMixIndA

##Plot
ggplot(data = graphdf1genderFeelProgPrepMixIndAcad) +
  geom_bar(mapping = aes(x = Gender, fill = FeelProgPrepMixIndAcad), position = "fill")+
  scale_colour_brewer(palette = "BrBg") +
  scale_fill_brewer(palette = "BrBG") +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

```
## Warning in pal_name(palette, type): Unknown palette BrBg
```



Appendix: Misc Example Code Area

How to rename once

```
###AV Script to rename one or multiples, same for multiples but just repeat first line till the end and

#df14 <- names(df13)[1] <- "Collecting"
#df14 <- df13
```

Basic plot

```
#ggplot(data = df) +
    ###actual plot bit
 # geom_bar(mapping = aes(x = Gender, fill = Collecting), position = "fill")+
    ###color
  #scale_colour_brewer(palette = "BrBg") +
   #scale_fill_brewer(palette = "BrBG") +
  ### Makes sure that the names/titles/labels do not overlap
     #scale_x_discrete(guide = guide_axis(n.dodge=2))
```

Creating new data frames by selecting only rows with a certain value

```
#Possibly create new dataframes by row value
#df12Male <- df12[df12$'Gender' == 'Male (including trans male)',]
#df12Female <- df12[df12$'Gender' == 'Female (including trans female)',]
#df12GenderQueerandorNonBinary <- df12[df12$'Gender' == 'Gender queer / Non binary',]
```

This will be the bread and butter of this assignment. It looks beautiful. Waffle Graph, must have info input manually. Still beautiful.

```
#parts <- c(`Un-breached\nUS Population` = (318 - 11 - 79), `Premera` = 11, `Anthem` = 79)

#waffle(
 # parts, rows = 8, size = 1,
  #colors = c("#969696", "#1879bf", "#009bda"), legend_pos = "bottom"
```

Appendix: Misc Example Code Area

How to rename once

```
###AV Script to rename one or multiples, same for multiples but just repeat first line till the end and

#df14 <- names(df13)[1] <- "Collecting"
#df14 <- df13
```

Basic plot

```
#ggplot(data = df) +
    ###actual plot bit
 # geom_bar(mapping = aes(x = Gender, fill = Collecting), position = "fill")+
    ###color
  #scale_colour_brewer(palette = "BrBg") +
   #scale_fill_brewer(palette = "BrBG") +
  ### Makes sure that the names/titles/labels do not overlap
     #scale_x_discrete(guide = guide_axis(n.dodge=2))
```

Creating new data frames by selecting only rows with a certain value

```
#Possibly create new dataframes by row value
#df12Male <- df12[df12$'Gender' == 'Male (including trans male)',]
#df12Female <- df12[df12$'Gender' == 'Female (including trans female)',]
#df12GenderQueerandorNonBinary <- df12[df12$'Gender' == 'Gender queer / Non binary',]
```

Analysis that isn't working yet

```
#lm1 <- lm(data = df, Collecting ~ Gender) # the model

#summary(lm1) # summarizes the output of the model
```

####Gender against Collecting Data first graph, not incredibly interesting and also hampered by small sample size of Gender Queer/Non Binary and Prefer Not To Say

```
#groupedBarImmCollecting <- ggplot(data = dfImm) +
 # geom_bar(mapping = aes(x = Immigration, fill = Collecting), position = "dodge")+
#  scale_colour_brewer(palette = "BrBG") +
 # scale_fill_brewer(palette = "BrBG") +
#  scale_x_discrete(guide = guide_axis(n.dodge=2))


#groupedBarGenderCollectingImm
```

A better proportional graph. This chunk also contains the code to reorder variables manually.

```
#data_newImm <- dfImm
#dfImm <- data_newImm

#dfImm$Collecting <- factor(dfImm$Collecting , levels=c("Very badly", "Badly", "Neither well nor badly"

#data_newImm <- dfImm
#dfImm <- data_newImm


#ggplot(data = data_newImm) +
#  geom_bar(mapping = aes(x = Immigration, fill = Collecting), position = "fill")+
#  scale_colour_brewer(palette = "BrBG") +
#  scale_fill_brewer(palette = "BrBG") +
#  scale_x_discrete(guide = guide_axis(n.dodge=2))
```

Analysis for Collecting Data Variable against Gender, Count and Proportion

```
###Output A Table and dataframe, possible first step in analysis given grouping. difficult to say, woul
#Collecting <- table(df12['Collecting'])
#genderVsCollecting <- as.data.frame.array(Collecting)
#genderVsCollecting


###Proportional Analysis
#CollectingProportion <- as.data.frame(table(df12$Collecting)/length(df12$Collecting))
###Display Output
#CollectingProportion
```