# SEIS631 Final Project: Big Five OCEAN Personality Analysis

Nikolai

5/4/2022

## Introduction

Personality has always been a fascination of mine. While many personality tests have little validity or predictive value, The Big Five personality tests have a reputation of being a good measure with predictive power. I have found a publicly available dataset of demographic information and personality data that I wish to analyze using techniques taught in class.

### What

As mentioned earlier, the Big Five Personality test is thought to be one of the more accurate and predictive personality tests for human behavior. For example, having a combination of high Agreeableness and Conscientiousness usually means the person is going to be a good employee while having high neuroticism is related to a variety of bad outcomes. I would like to see if there are relationships between the expression of different personality traits across countries and gender.

This dataset was not created by myself, and can be found at Kaggle

### Why

While we know that the expression of certain combinations of personality traits predict certain behaviors, these traits are often seen as being separate and distinct measures of personality. I argue that there are likely 'personality types' - combinations of these traits that are more likely to occur together - that define subsets of the population.

### How

My plan is to subset the dataset while controlling for the expression of one or more personality traits. Using these traits as a control, I will look to see if the other traits expression in this subset is significantly different from the normal population.

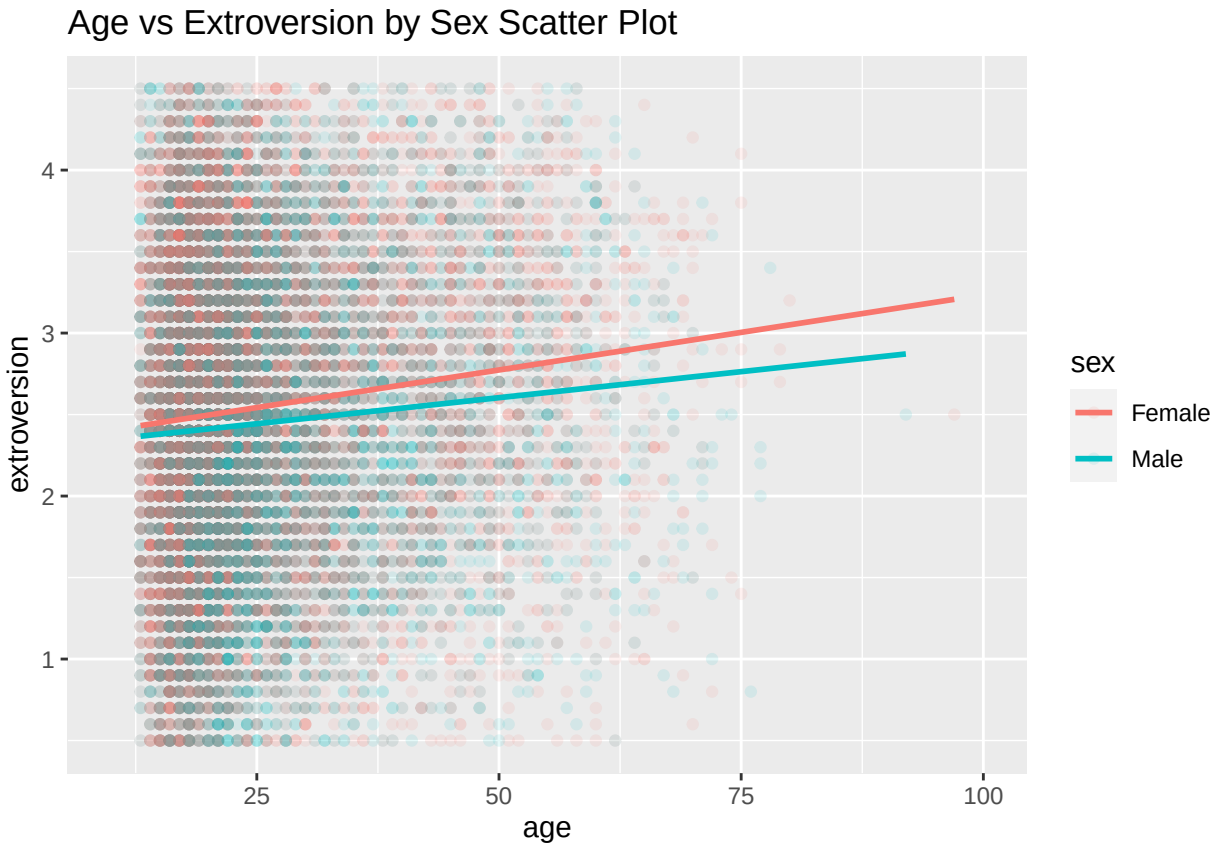
## Body

### Cleaning the data.

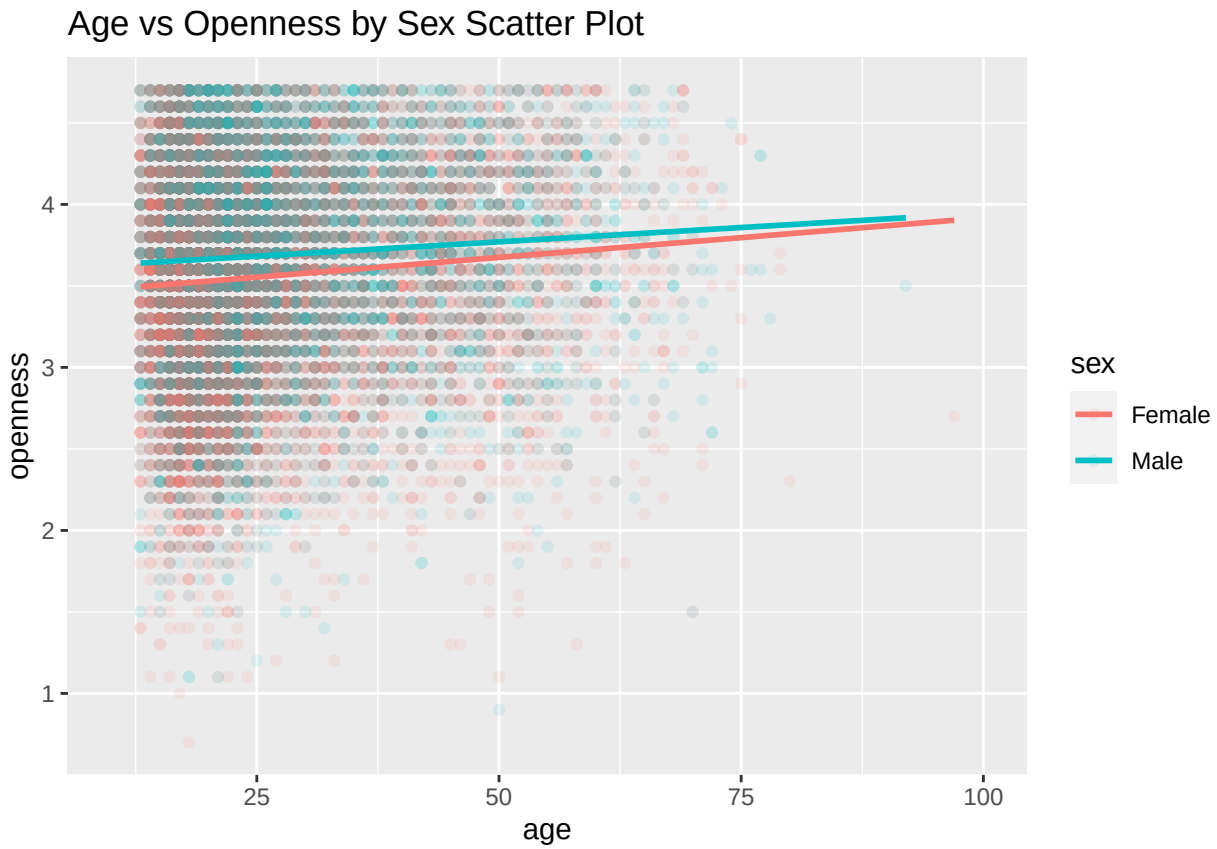
The dataset was quite large and had demographic information and the actual responses to the individual questions. While this is ideal, it means the dataset must be cleaned to be useful. For example, to account for the tendency to give high ratings over low ratings regardless of the question's content, several of the questions are stated negatively. For example, on a rating of 1-5, answering the question 'I like to talk' with a 5 would be high extroversion, but answering the question 'I don't like loud parties' with a 5 would actually be the opposite. I will write a function to inverse a question's rating and then manually code which questions are positive or negative and apply the function only to negative questions. Then, because there would be too many rows with this dataset if looking only at the individual questions, I will average the answers per trait to get a single score for the five traits.

**Viewing the raw data to analyze major trends in age and gender**

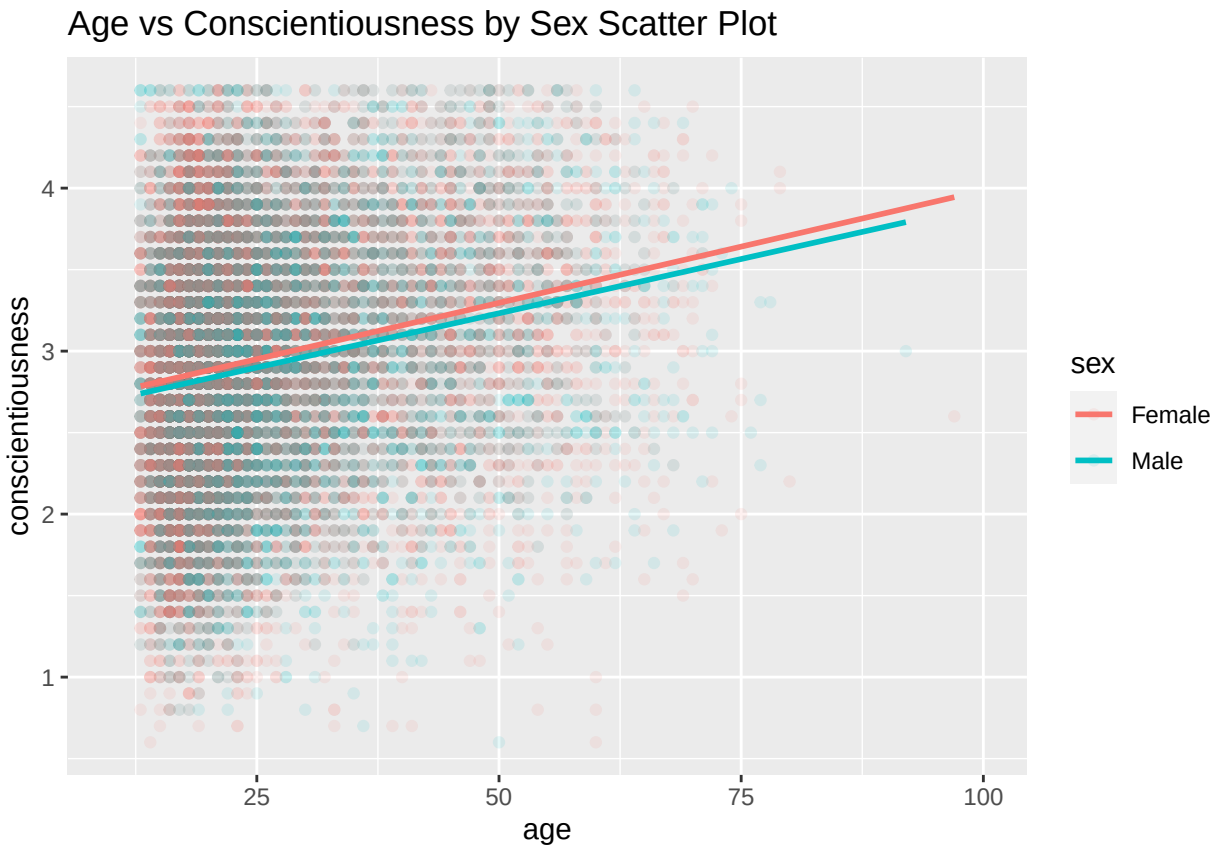
After cleaning the data, I decided to plot the data using scatter plots comparing each trait to age and gender to see if I could see any glaring trends.



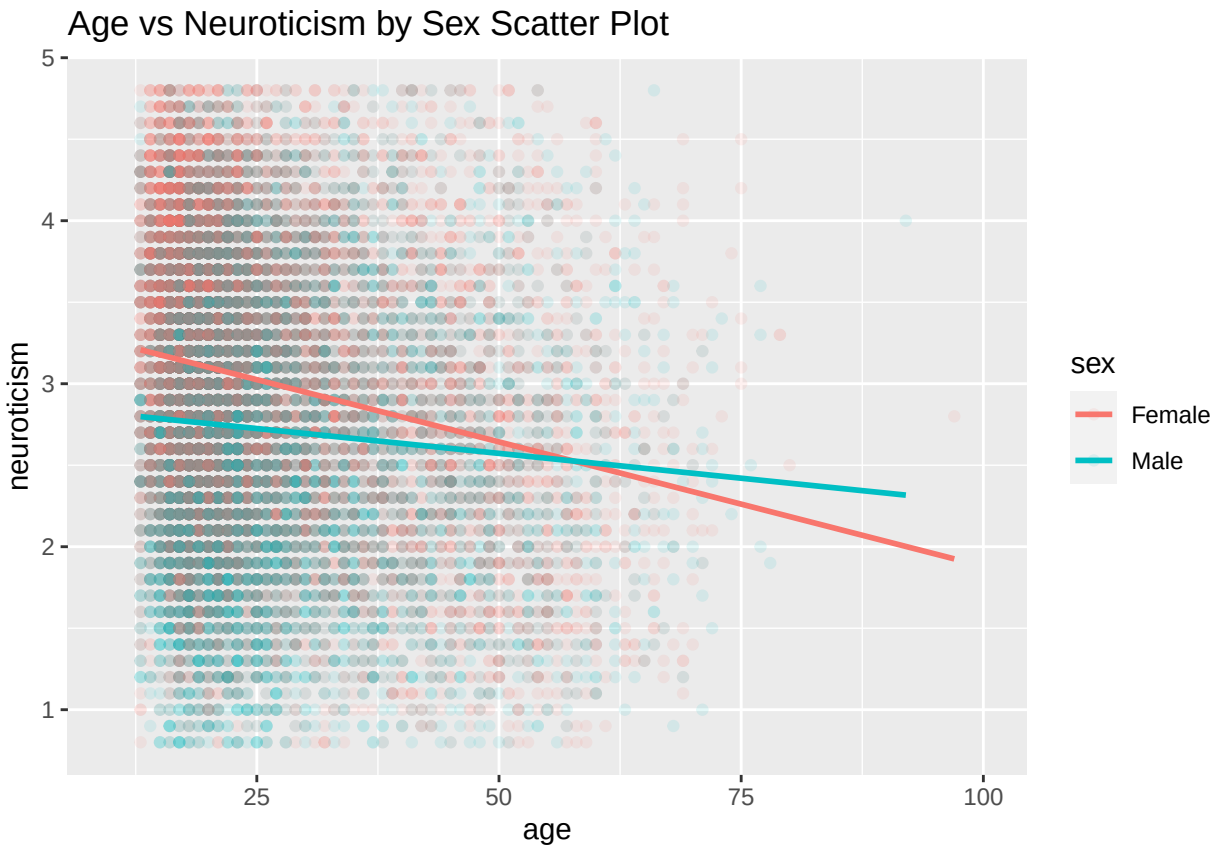

Extroversion had a great degree of variability, ranging all the way to the extremes of 5 and 0. There does appear to be a trend towards older respondents being more extroverted, and this trend seems more prominent in the female population, but with this kind of variability I would not expect the correlation to be very high, or the trend to be very significant or explanatory.



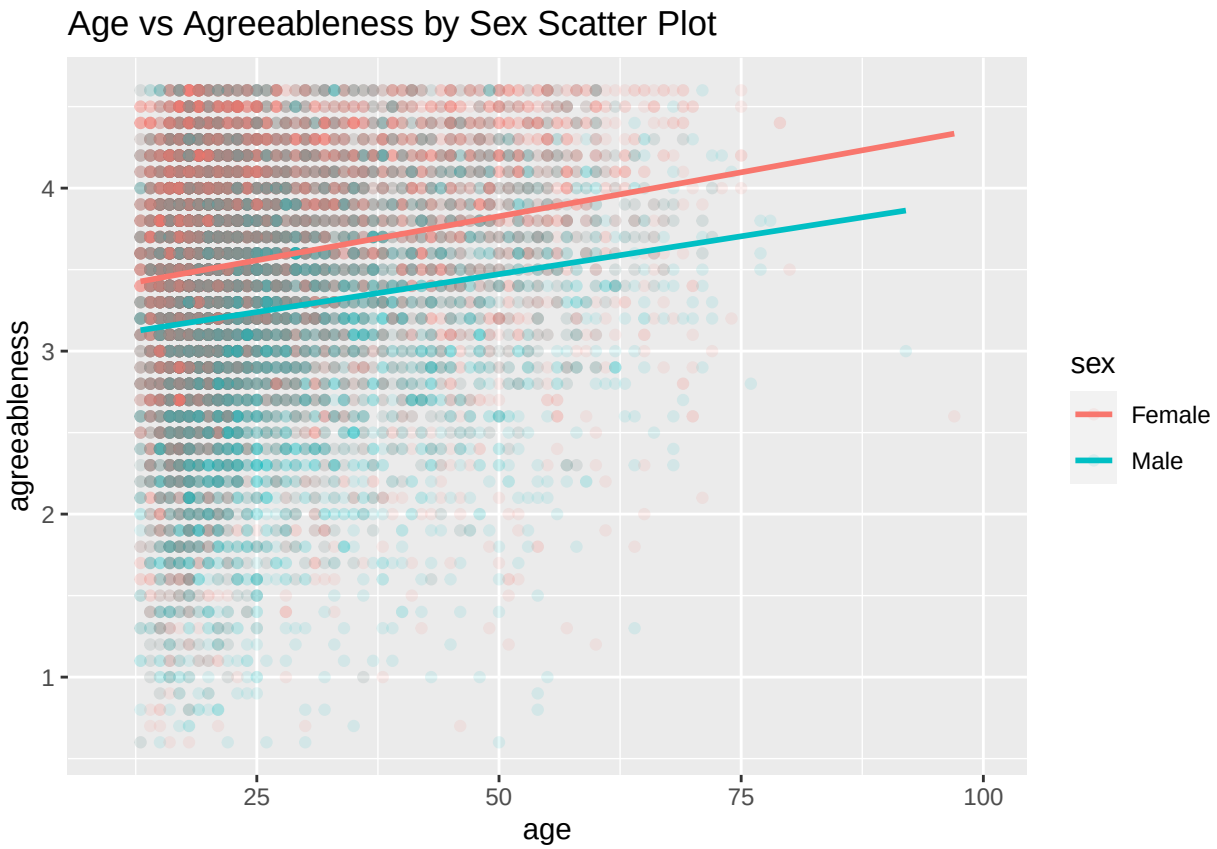

The data seems to indicate with openness to experience that males tend to rate themselves higher than females slightly. I can see from the data here that most people, most of the time rate themselves high on openness. This is a little disappointing, because it means that openness will not be a good explanatory variable. I am expecting to see little significant differences between populations regarding openness.




Conscientiousness does not appear to have significant differences between male and female populations, but does appear to have an upward trend regarding age. The spread of the points, however, is large and I am expecting the correlation to be small. Older populations seem to rate themselves consistently higher than a 2 on conscientiousness.

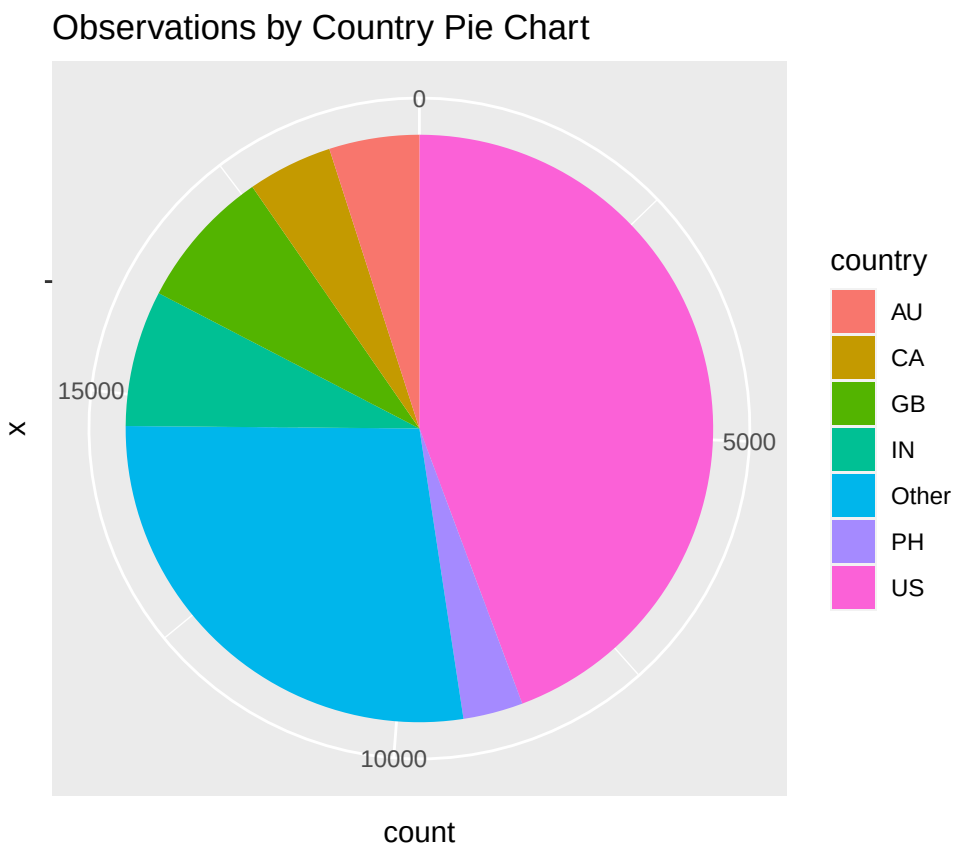



Neuroticism appears to be very interesting. Not only do I see a negative trend in the data over age, but also the regression lines between males and females have very different slopes and actually cross in the data. It appears like males have a greater spread than females in responses, resulting in a very level regression line. Females likely rate themselves as more neurotic in younger populations and less neurotic in older populations than their male counterparts.



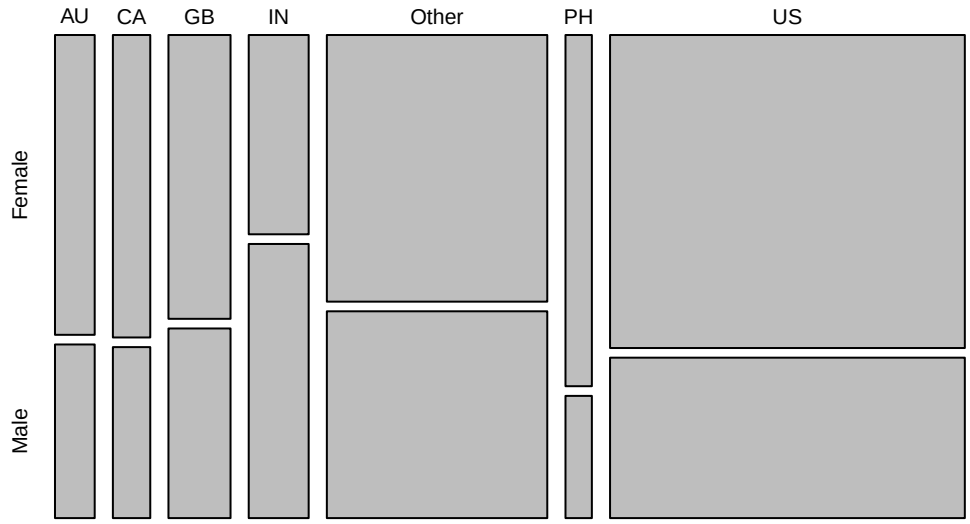

Agreeableness also is interesting and likely has significant differences. The regression line and scatter plot show that as age increases, so does agreeableness. The spread is also lower than we saw in the other traits. The regression line shows that while the slope of the regression line is similar for males and females over age, females appear to rate themselves as more agreeable than males consistently. I am expecting there to be significant differences in populations between these two groups.

**Trends in Country, Gender, and English Native Speakers**

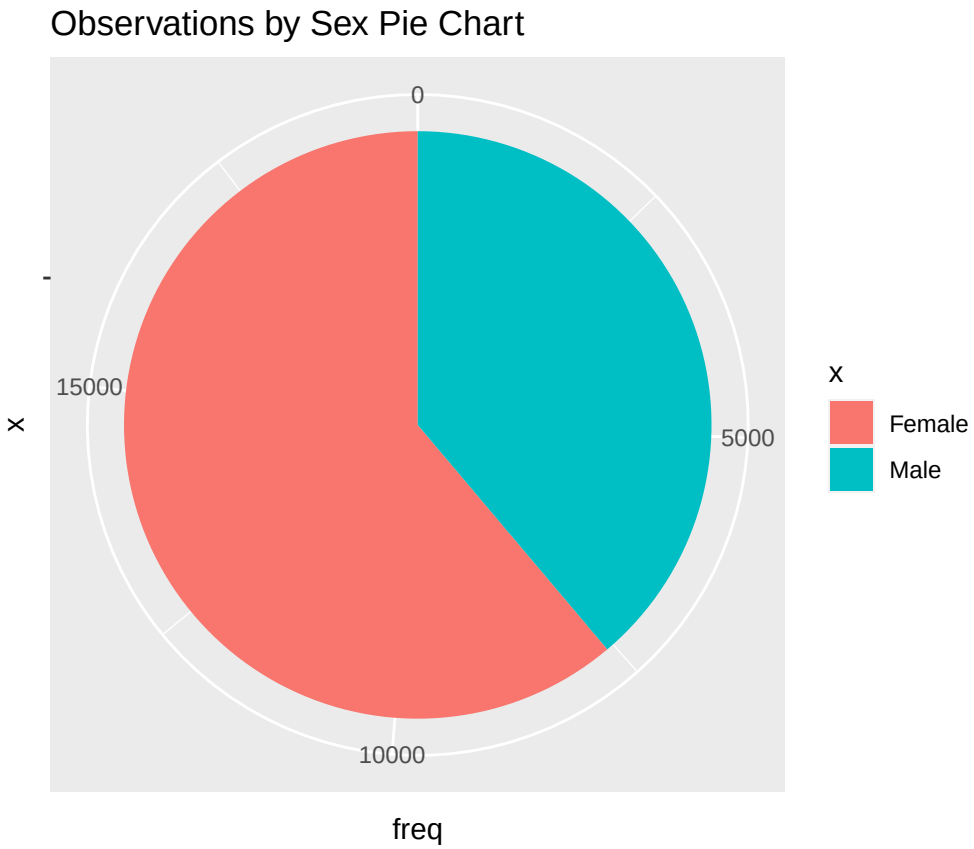



As seen by this R code, the top countries with observations in this dataset are English speaking countries. I will be looking at these countries next and similarities and differences within these countries to see if country, age, and sex result in significant differences. Here is a table that shows the observations per sex and country:

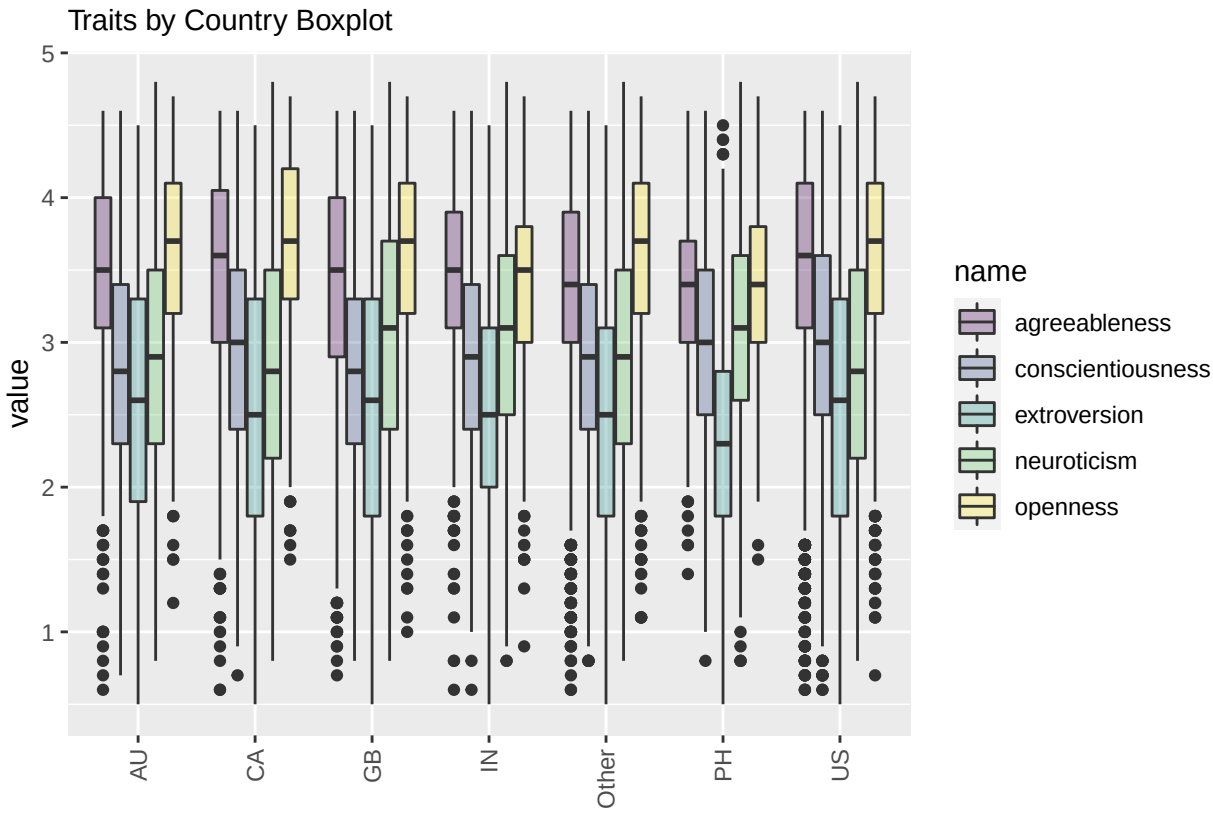
```
##
##         Female Male
##   AU       613  355
##   CA       582  329
##   GB       907  606
##   IN       615  846
##   Other   3026 2347
##   PH       477  166
##   US      5717 2930
```



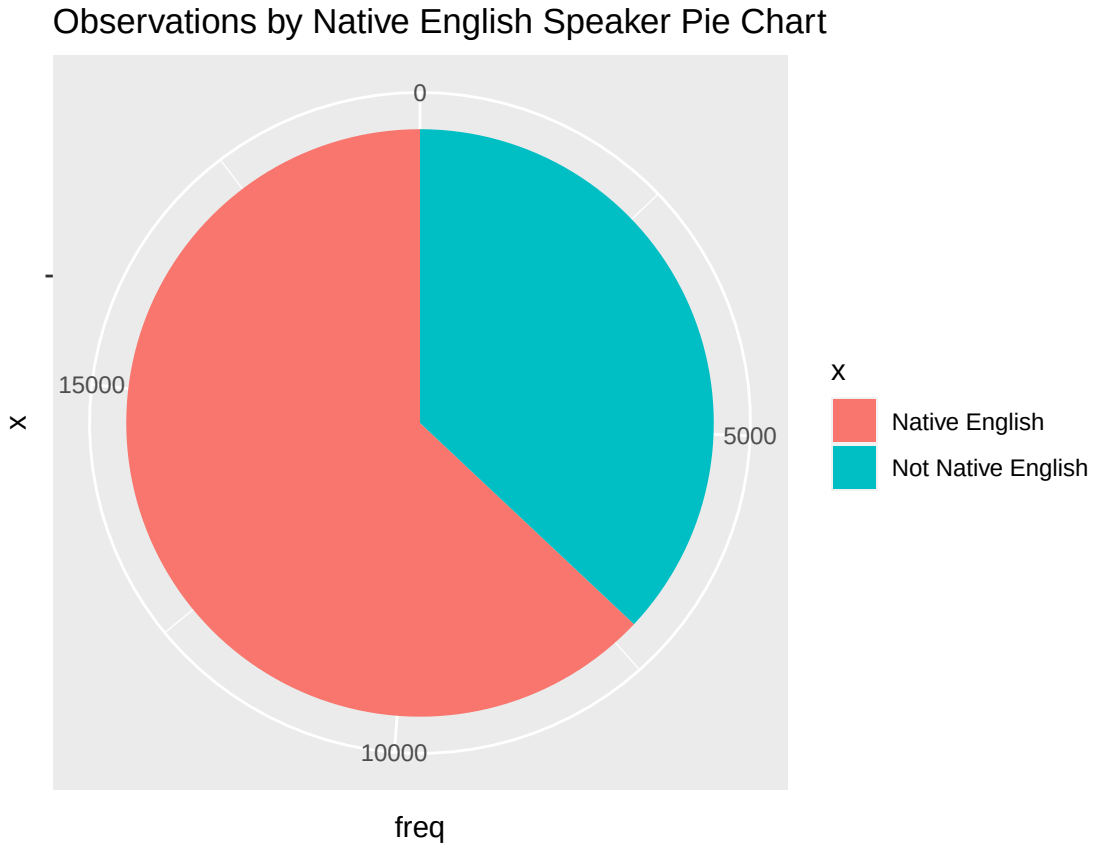

According to the data, we have a higher percentage of female respondents than males in all countries except India. Almost half of our respondents are from the US, and about 3/4 are from English speaking countries. This is a large dataset, so I'm not too concerned with fewer responses leading to less power, but from now on when doing statistics I will be using density functions to determine the percentage of respondents versus the actual counts as the counts of the comparing populations vary.



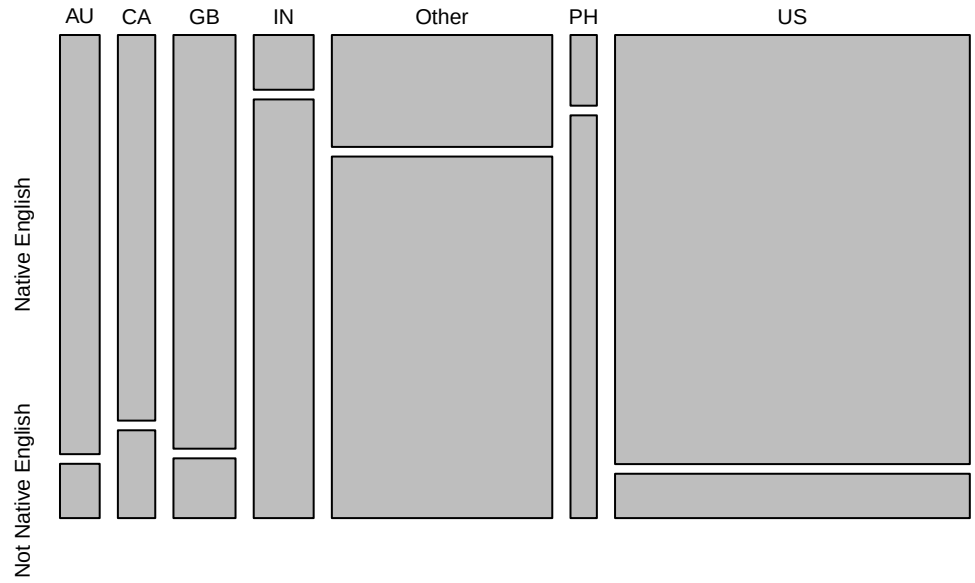

In fact, as seen in this pie chart, we have approximately 2/3 females to 1/3 male respondents overall. This again illustrates the need to use density equations and normalized graphs when analyzing the data.




Looking at the boxplot of the different countries and their responses, what strikes me is we have a very similar distribution of personality across countries. We are seeing the same shape and standard deviations- that being high agreeableness and openness, medium conscientiousness and neuroticism, and a tendency towards low extroversion with high variability. My expectation after seeing these graphs is that there will not be many, if any, significant or interesting differences between countries.



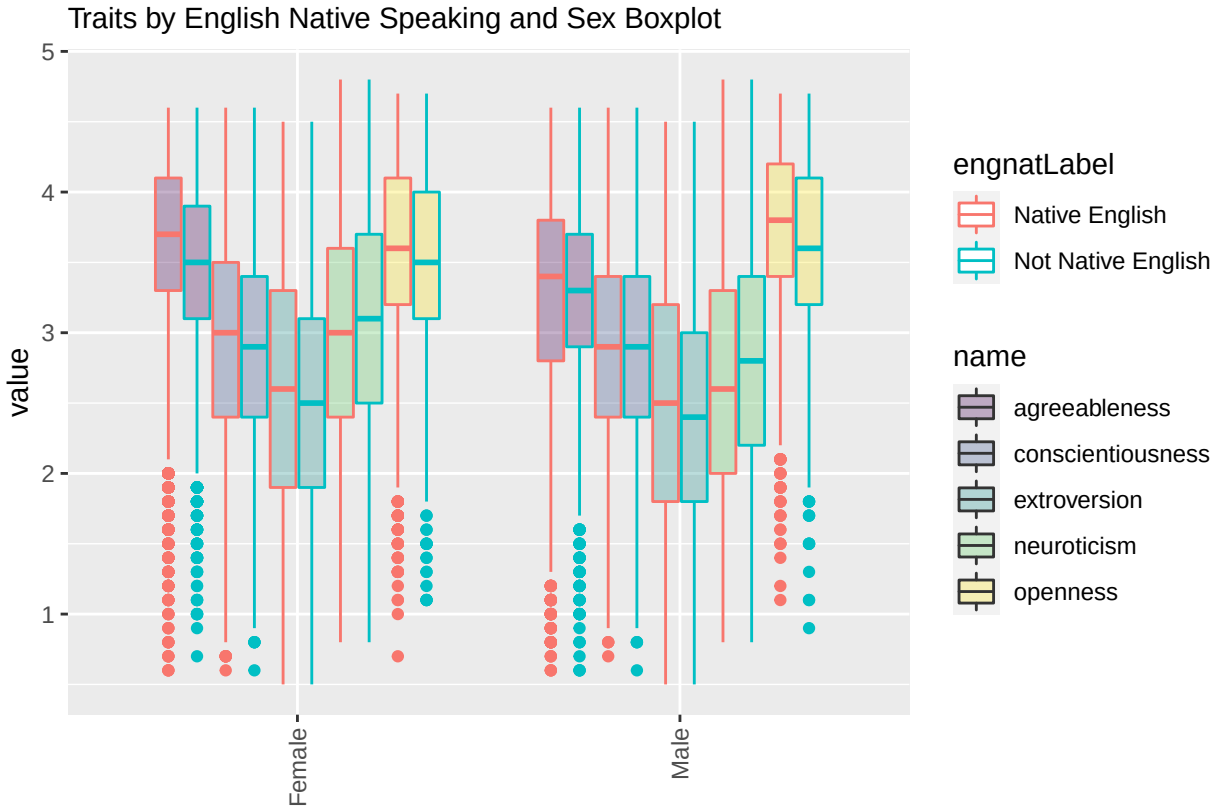

Another interesting statistic of this data is that, although 3/4 of the respondents are from English speaking countries, only about 2/3 of the respondents are native English speakers.



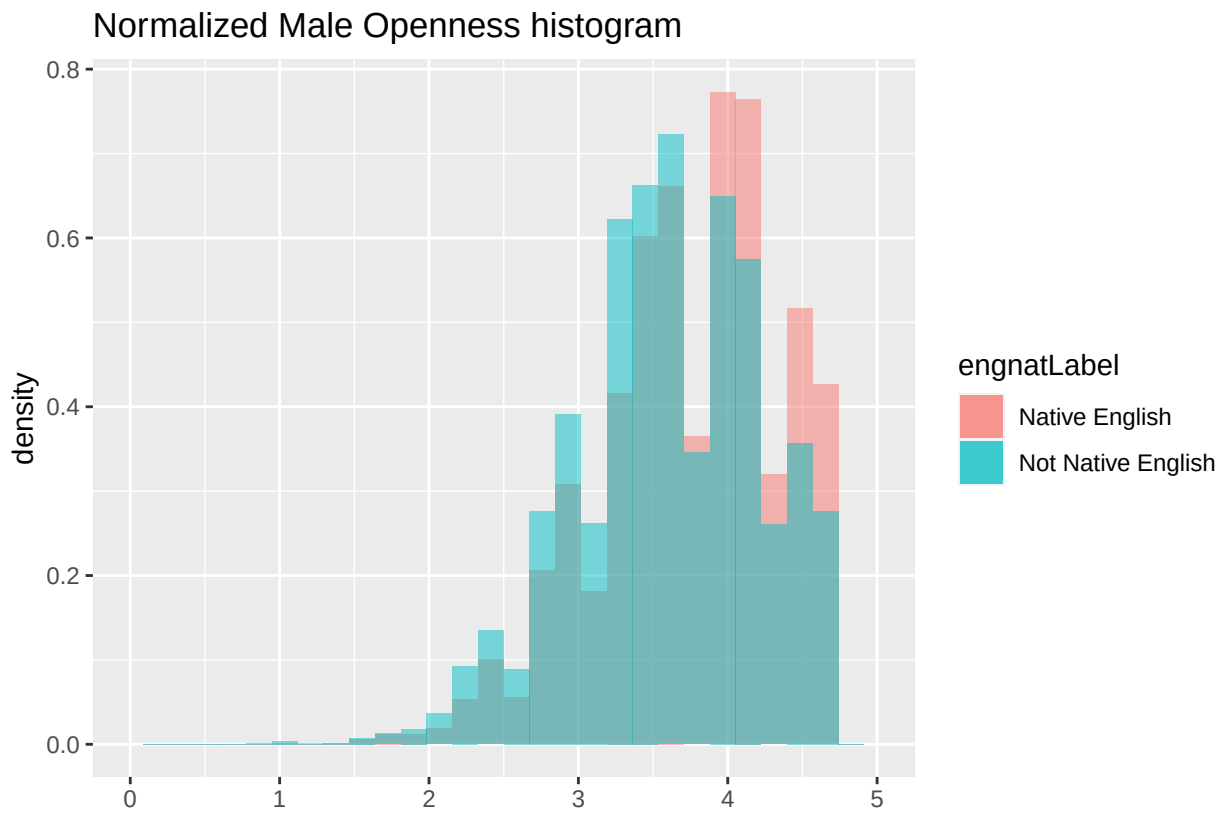

```
##
##         Native English Not Native English
##   AU               857                111
##   CA               742                169
##   GB              1322                191
##   IN               169               1292
##   Other           1270               4103
##   PH                96                547
##   US              7837                810
```

Which, according to this mosaic plot, those non-native speakers mostly come from India and the Philippines. Which makes sense given their history and language data compared to other English speaking countries. Due to the fact that the Big Five is an English based test, we can expect most respondents to be English speakers. Also, because it's an American test, it is not surprising that the popularity of the test focuses on English speaking countries.

**Trait Trends in Sex and English Nativity**

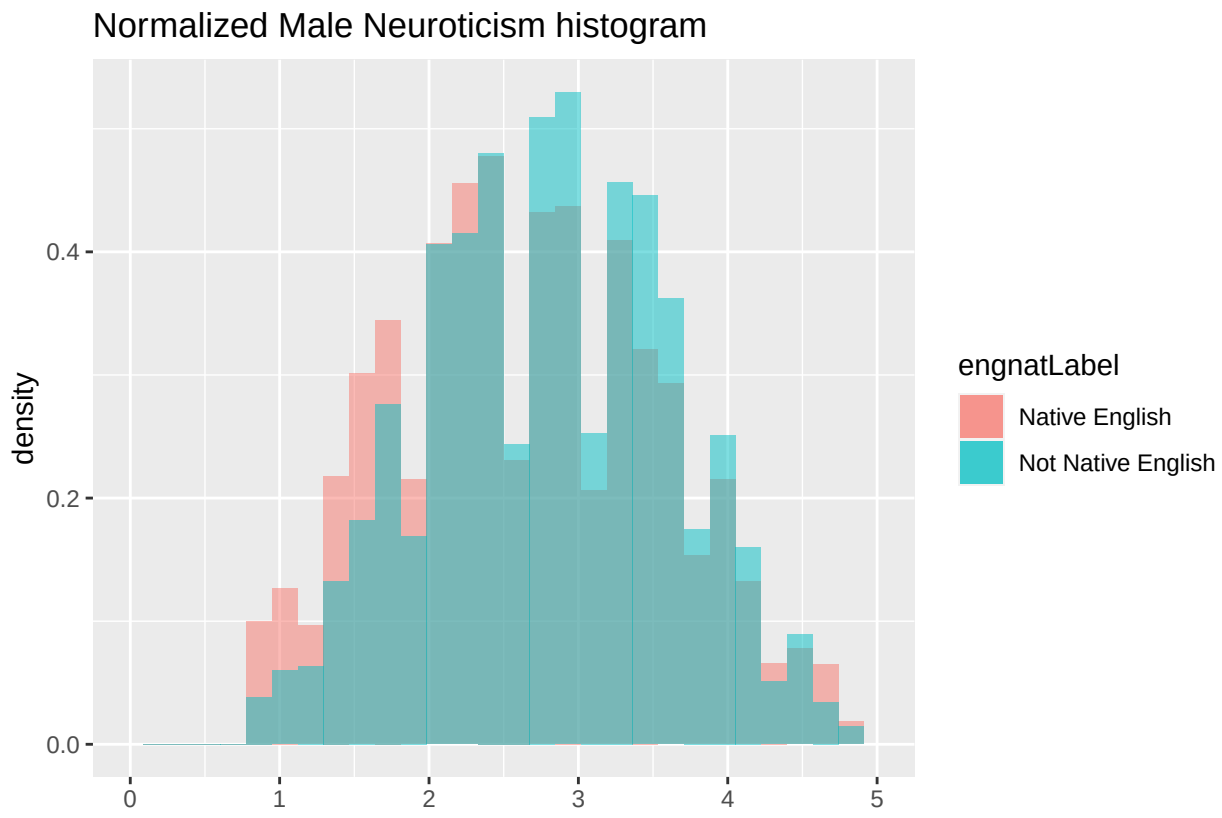



Looking at the differences between the native English and non native English populations and breaking it down by sex, I noticed that while the female population did not seem to have big differences in trait expression with the possible exception of native speakers having higher agreeableness, the male population of native speakers tended to have higher openness and lower neuroticism. I decided to investigate these differences next.




```
##
##  Welch Two Sample t-test
##
## data:  ENmales$openness and NENmales$openness
## t = 10.986, df = 6758.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1270268 0.1822052
## sample estimates:
## mean of x mean of y
##  3.752231  3.597615
```

Looking at the normalized histogram between native and non-native English speaking males, I could see the native population had a tendency to rate themselves higher in openness, as expected. I ran a t-test, and the p-value is lower than .05, indicating that the difference in means is a significant difference. It is, however, not a big difference, being 3.75 for English native males and 3.60 for non-native males.

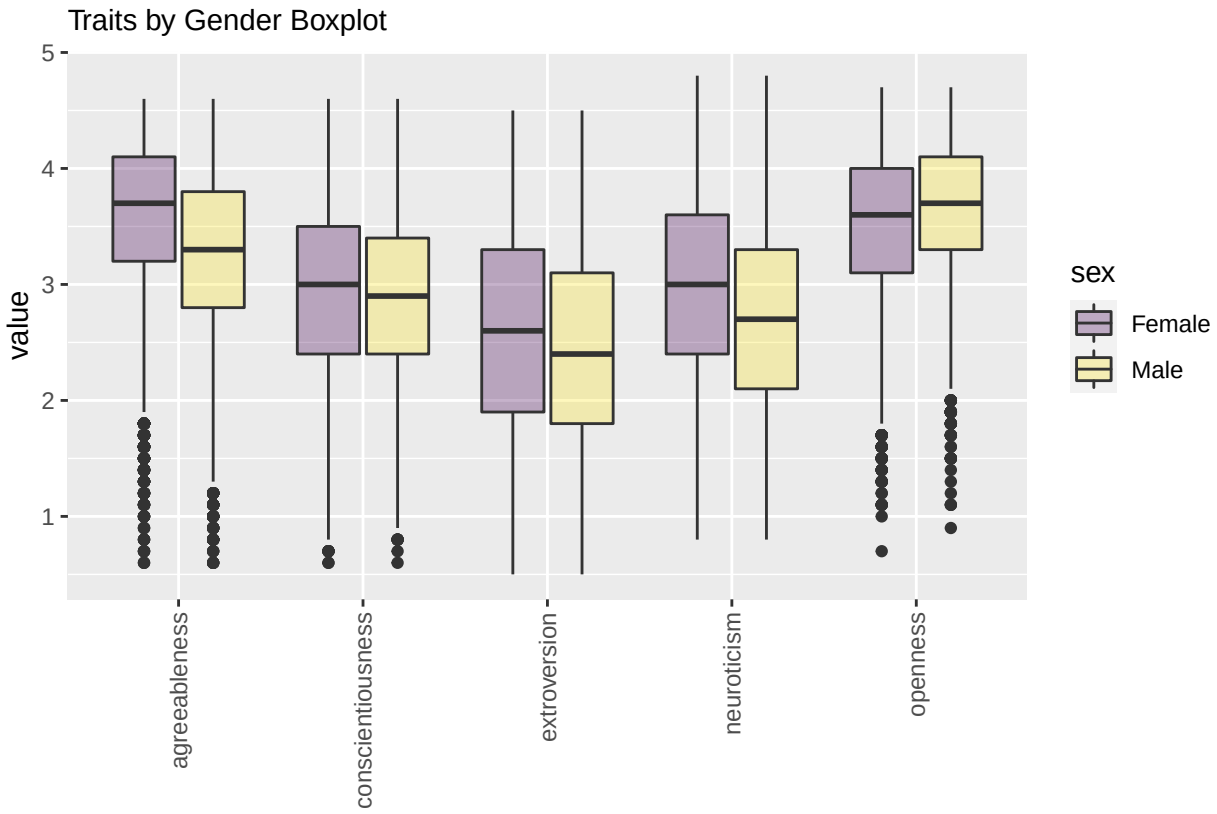



```
##
##  Welch Two Sample t-test
##
## data:  ENmales$neuroticism and NENmales$neuroticism
## t = -7.9124, df = 7189.6, p-value = 2.903e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1943073 -0.1171452
## sample estimates:
## mean of x mean of y
##  2.649294  2.805020
```

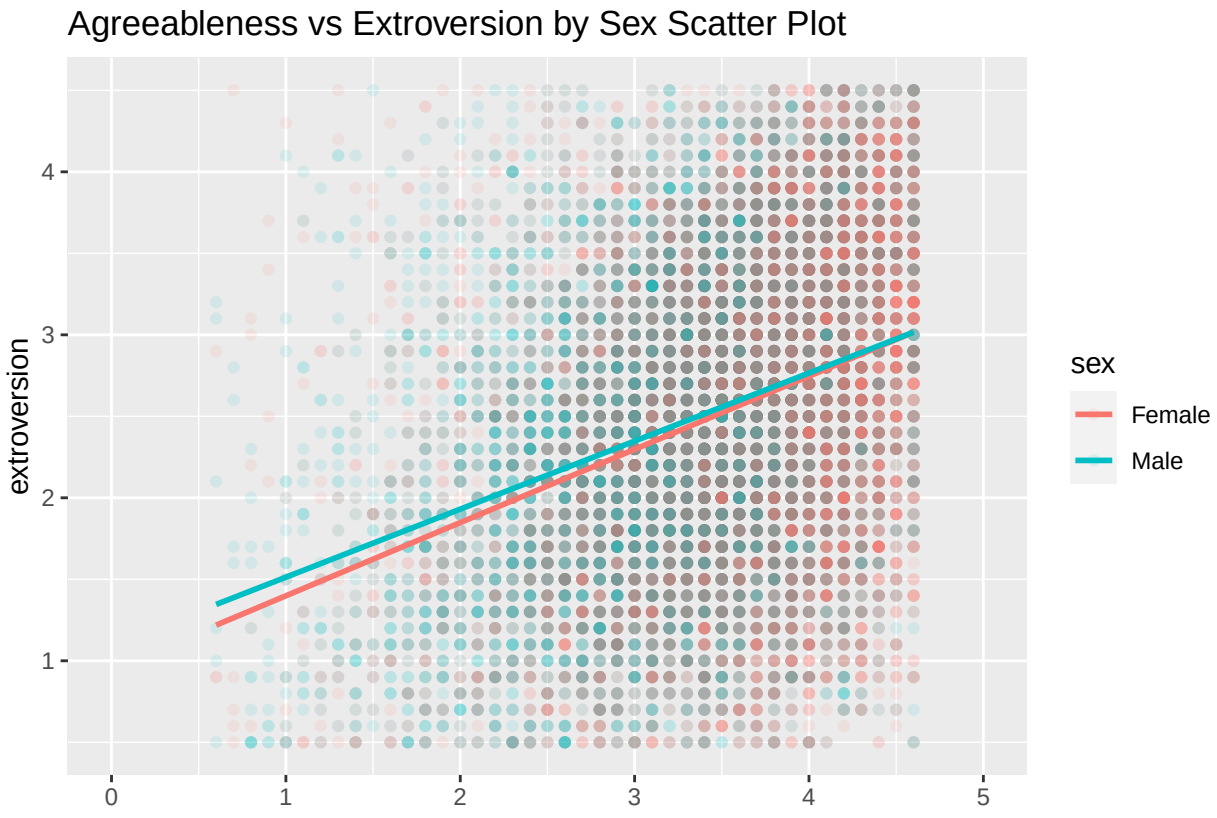
I found similar results with neuroticism, except this time English native males scored lower on neuroticism than their counterparts. This also represented a significant difference, with a greater difference in means- 2.65 for native and 2.81 for non-native.

### Correlations between Traits and Gender

After investigating the effects of being an English Native Speaker, I became interested in delving into the effects of traits on each other and sex. When looking at the boxplot of sex and the traits, I noticed the female population seemed to be more agreeable, neurotic, and extroverted than the male population. Conscientiousness, on the other hand, did not appear to have significant differences.

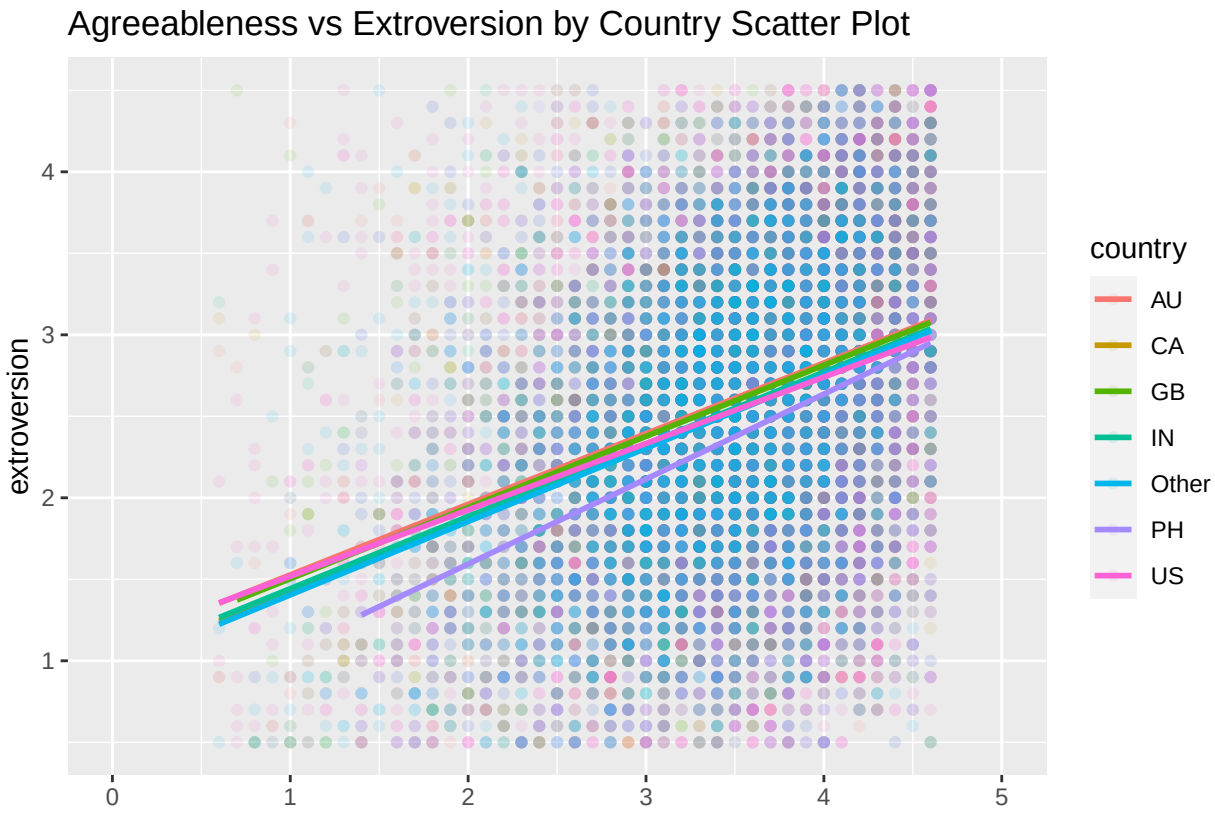



This gave me the idea to plot the relationship between the traits themselves and look for differences in gender expression.

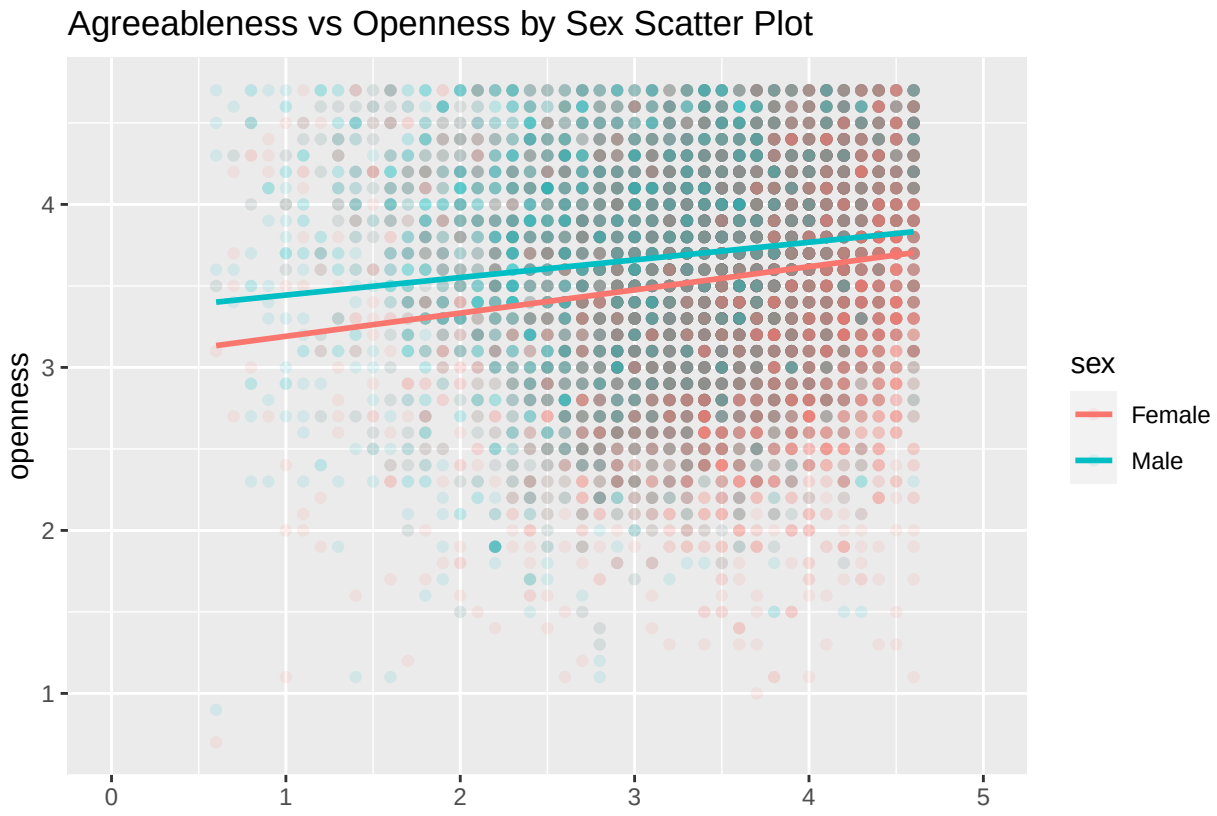



```
## [1] 0.333051
```

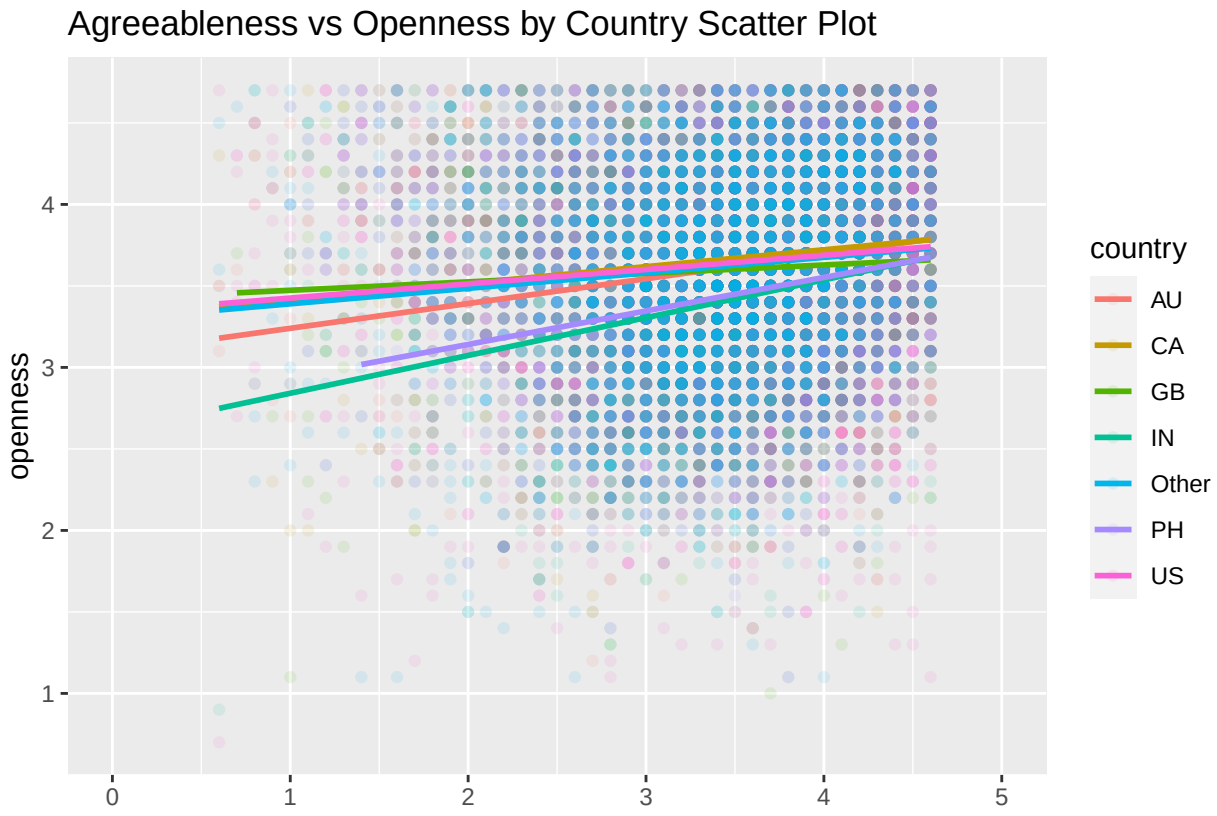
First I checked the relationship between agreeableness and extroversion. According to the graph, there is little difference between the male and female populations, though both had a positive correlation. Simply put, as extroversion increased, so did agreeableness. I checked the correlation coefficient, which turned out to be .333. This means that, while the correlation exists, it is a weak correlation and there is a lot of variability int he dataset. As a follow up question, I decided to look into the differences between countries on this relationship.




In accordance with the previous findings, most countries displayed a similar trend. However, I did find that the average low agreeable Filipino person had a trend towards being lower in extroversion than other countries. This makes me wonder, what about the Philippines makes their personality ratings so different?



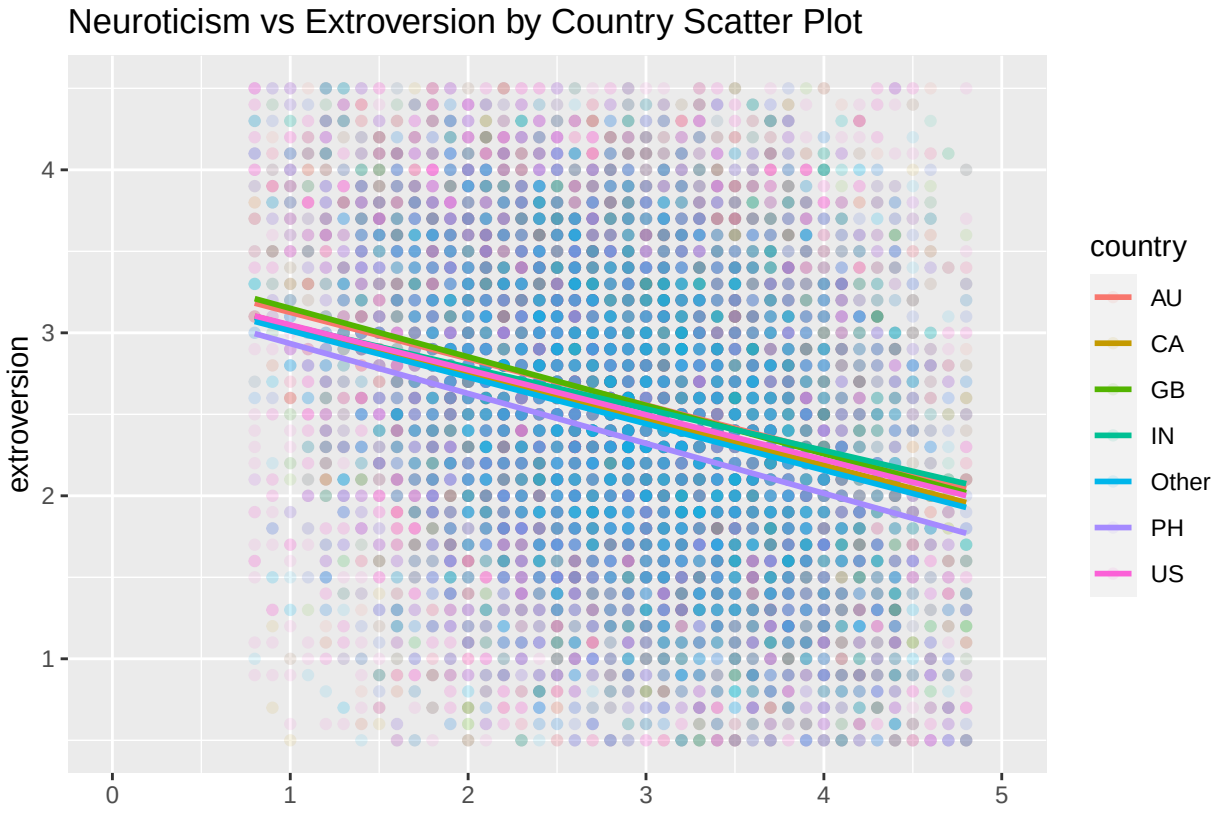

When looking at the relationship between agreeableness and openness, it seems there is a lightly correlated slightly inclined relationship. According to the data, males appear to be more open to experience than females in general. This also becomes interesting when looking at the same graph but with countries.



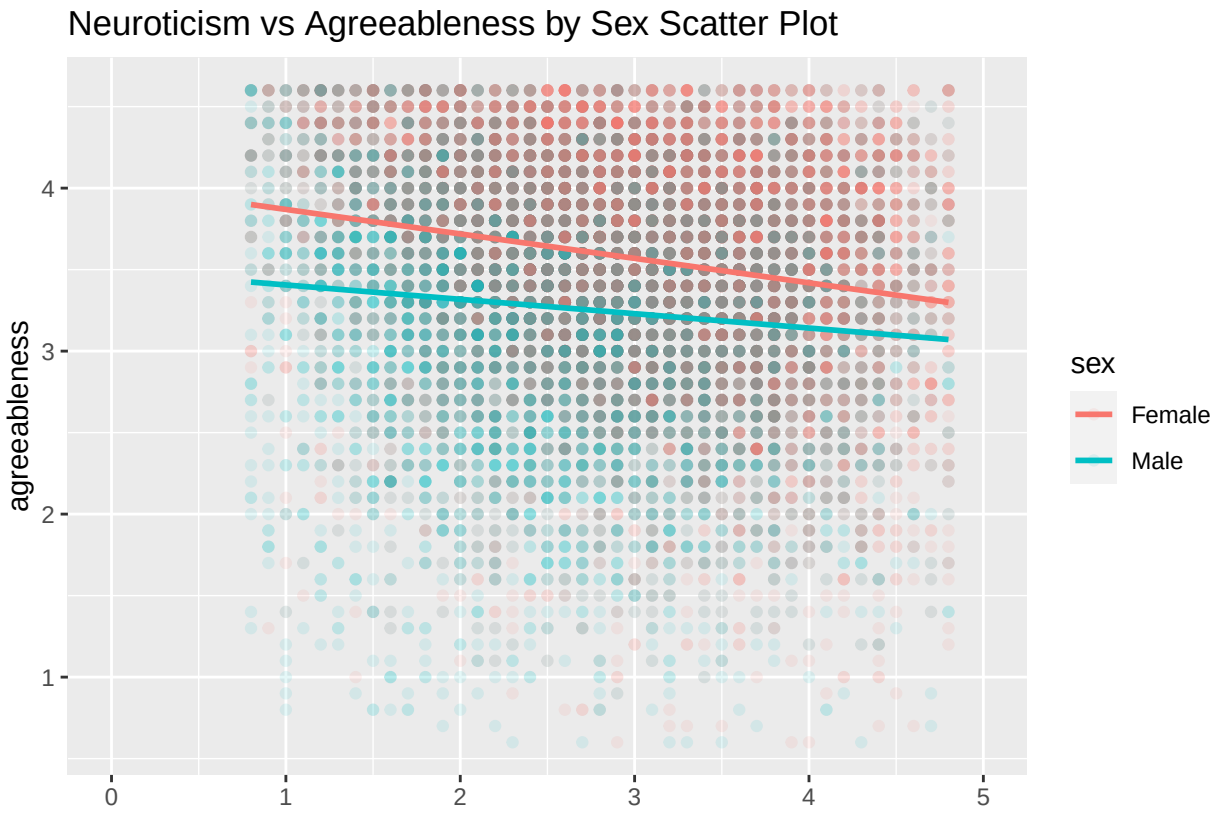

For an unknown reason, out of all of the countries measured including the average of 'Other' countries, the Philippines and India not only display differences with all other countries, but also display similar regression lines with each other. While at high agreeable levels their openness levels converge with the normal levels of other countries, at low levels they have less openness than normal. This could be a direction of further research.


```
## [1] -0.2624741
```

Looking at the relationship between neuroticism and extroversion, there appears to be the same negative correlation within male and female populations. It's also more apparent in this graph that there is a consistent difference in the population mean for extroversion- females tend to rate themselves higher than males. The correlation coefficient turned out to be -0.262, so it is present and negative, but only slightly correlated.

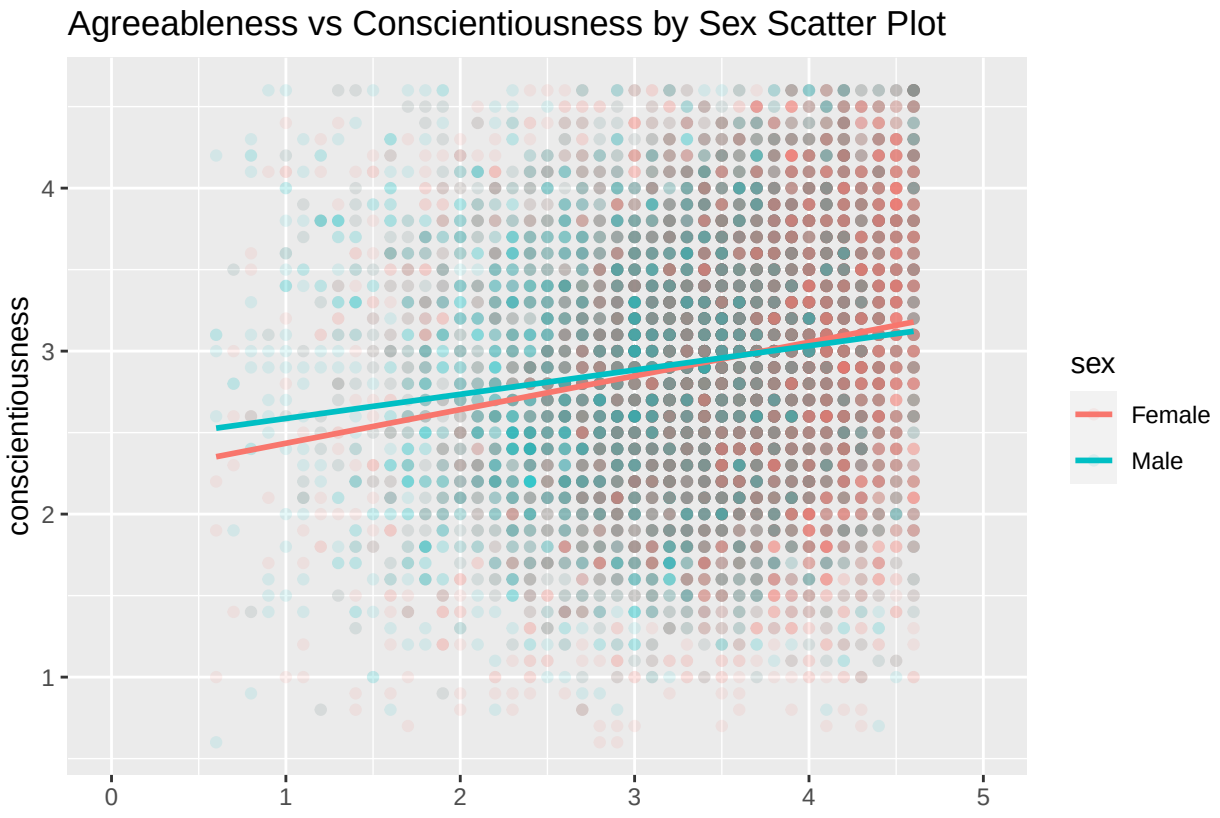



And, this trend appears to be present in all countries measured, even if it is slight.




```
## [1] -0.1117085
```

Looking at the relationship between neuroticism and agreeableness we see that females tend to rate themselves as more agreeable than males. There also appears to be a slightly negative correlation between agreeableness and neuroticism. Stated simply, the more neurotic you are, the less likely you are to be agreeable. The correlation coefficient is not high, however, resting at only -.11. This means that while the relationship exists, there is a high amount of variability and the relationship is not very explanatory. I checked the same graph by country, but there didn't appear to be significant differences between countries regarding this relationship.

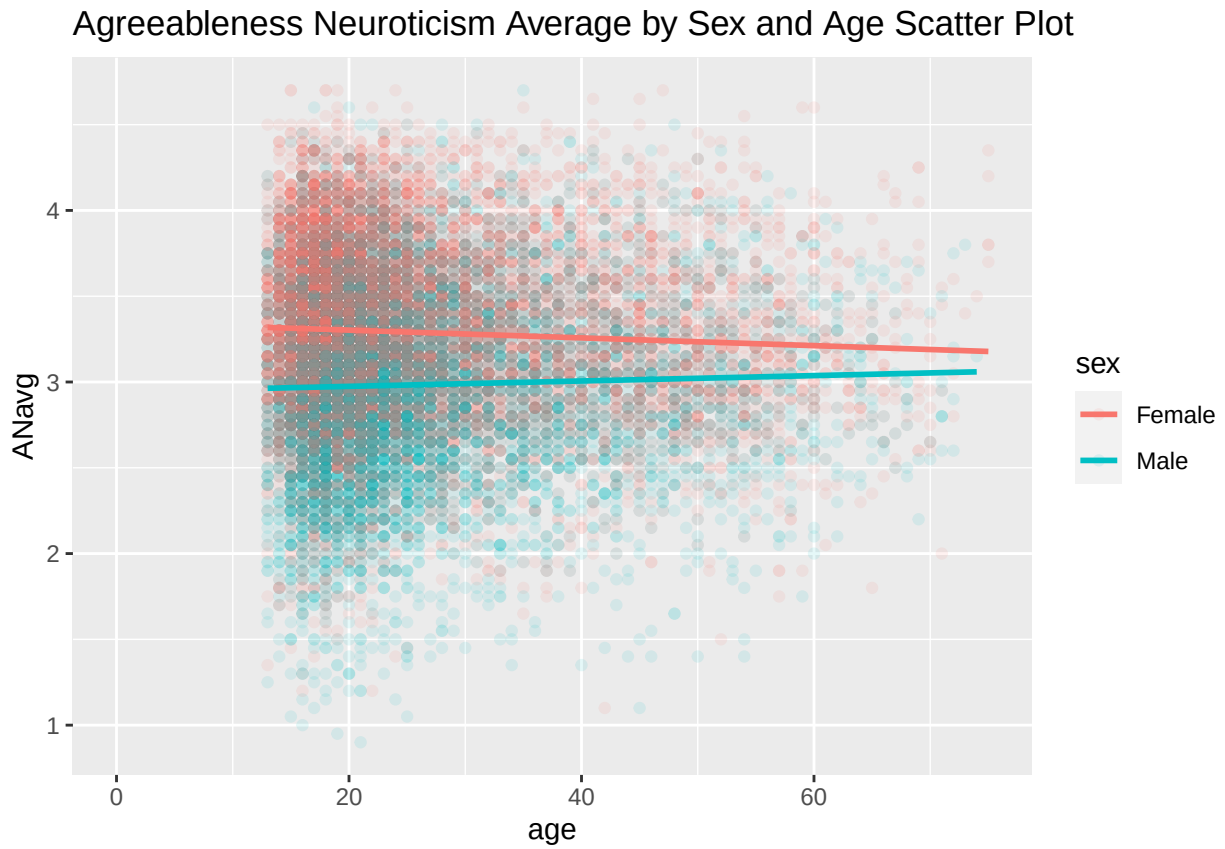



```
## [1] 0.1757425
```

And the last graph for this section is the relationship between conscientiousness and agreeableness. There is a slight correlation between the two, but like with other findings the relationship is not big. The correlation coefficient is only .176, which makes it larger than some of the relationships we've seen, but still rather small for explanatory purposes..

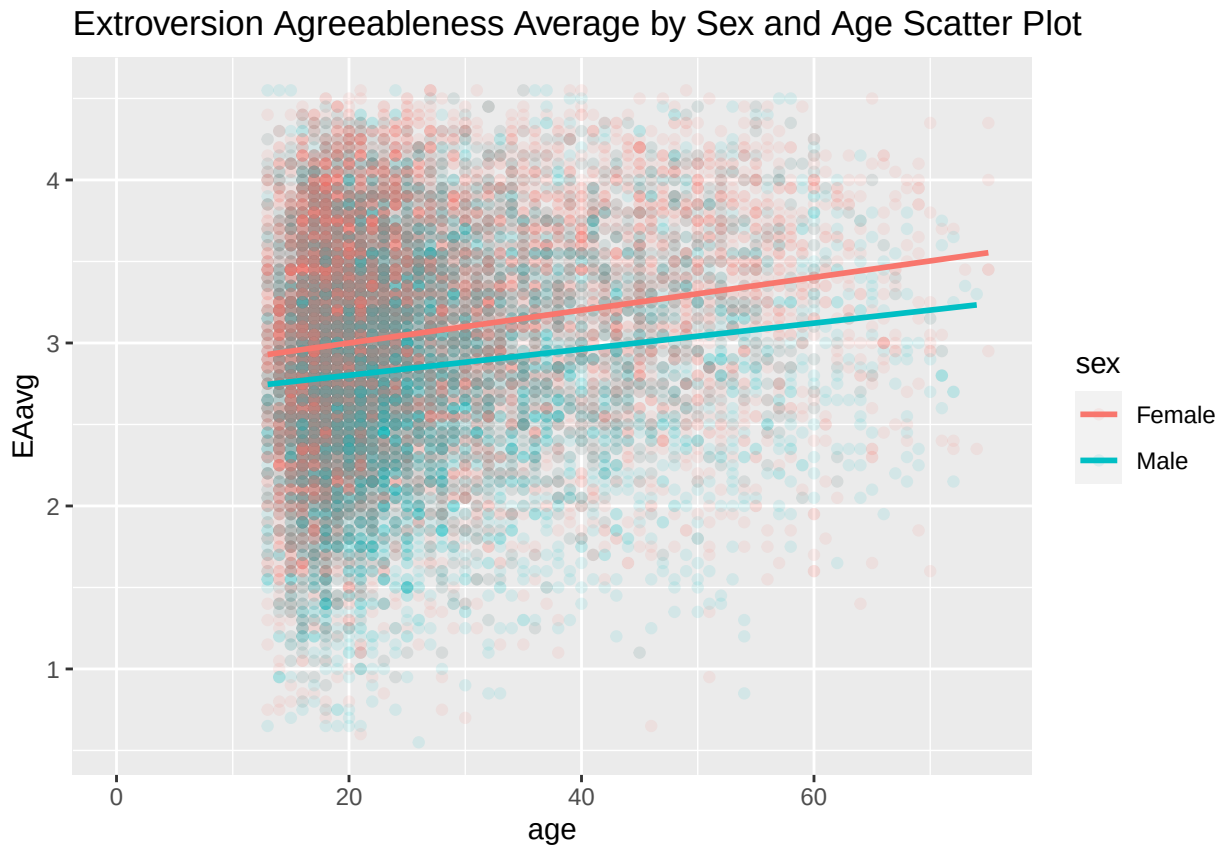
### Trait Combined Averages by Age, Sex, and Country

From the prior graphs, we know that there are several slight relationships between these traits and gender/country. I wanted to check if these relationships persist or change depending on population age.



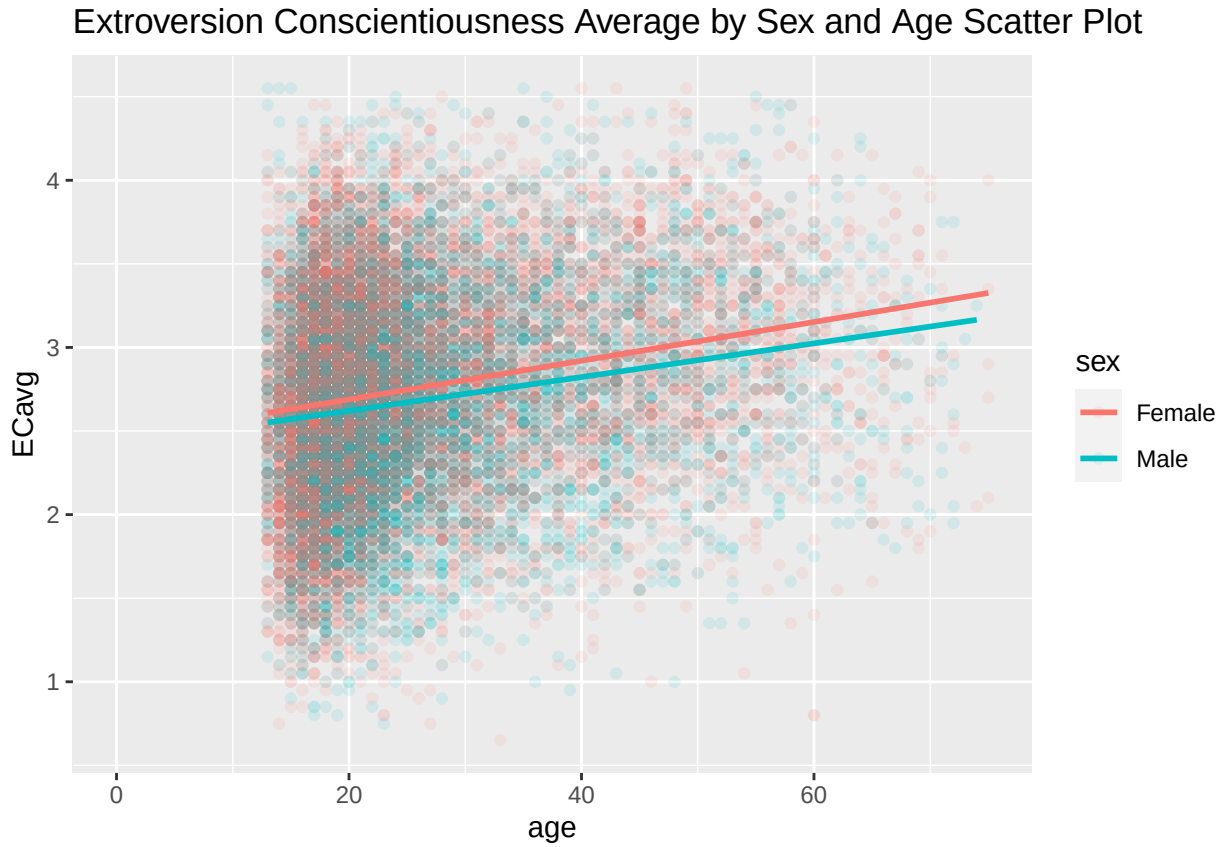

```
##
##  Welch Two Sample t-test
##
## data:  males$ANavg and females$ANavg
## t = -40.007, df = 15020, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3190131 -0.2892130
## sample estimates:
## mean of x mean of y
##  2.984833  3.288946
```

The first graph is the relationship between agreeableness and neuroticism. We already know that females tend to rate themselves as more agreeable and neurotic than their male counterparts on average. However, I wanted to check if this relationship changes based on age. The data indicates that, while females do tend to rate themselves higher on these two traits, the difference between males and females decreases as the population becomes older. I then ran a T test to determine if these populations were significantly different. The answer is yes, the relationship between agreeableness and neuroticism when looking at male and female populations is significantly different with a P value of under .05. The male population had a mean of 2.98 and the female population had a mean of 3.29, which appears to me to be a rather large difference.



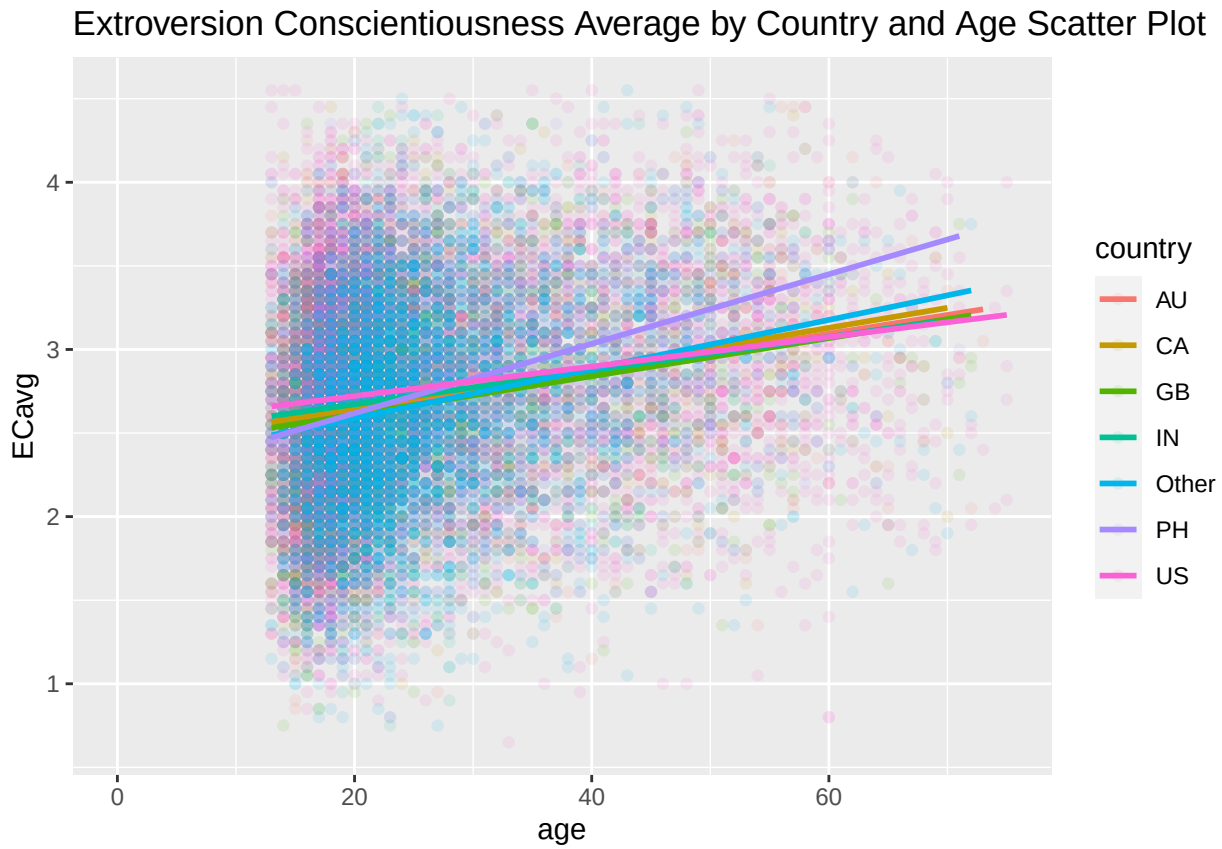

```
##
##  Welch Two Sample t-test
##
## data:  males$EAavg and females$EAavg
## t = -20.924, df = 15756, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2242907 -0.1858686
## sample estimates:
## mean of x mean of y
##   2.85473   3.05981
```

Looking at the relationship between extroversion and agreeableness, there also appears to be a consistent trend towards females rating themselves as more extroverted and agreeable than males. This relationship also appears to become slightly more pronounced with older populations. I ran a T test and found that this difference is also significant, having a P value of lower than .05. The means of each population are not as different as the means for agreeableness and neuroticism, however, with males having a mean of 2.85 and females having a mean of 3.06.

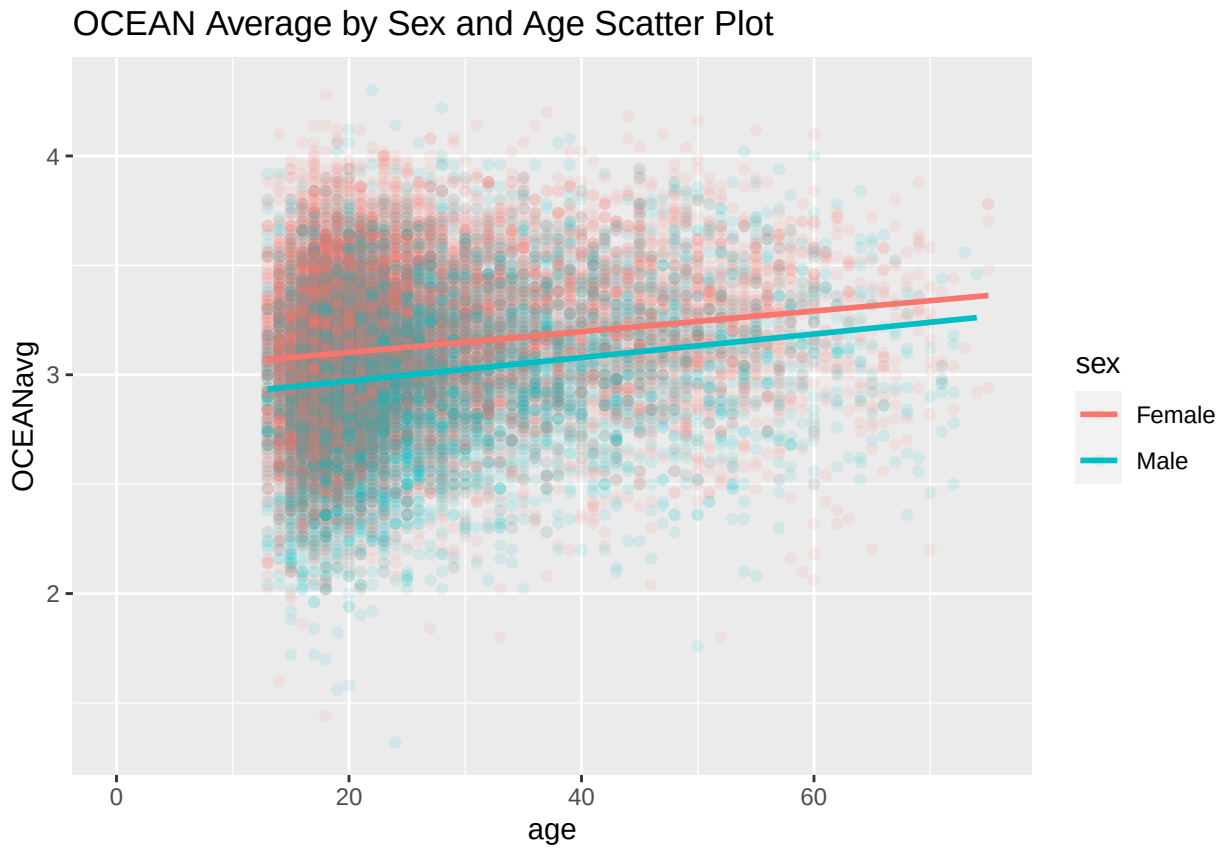



```
##
##  Welch Two Sample t-test
##
## data:  males$ECavg and females$ECavg
## t = -7.797, df = 16198, p-value = 6.722e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08823033 -0.05278117
## sample estimates:
## mean of x mean of y
##  2.688211  2.758717
```

Looking at the relationship between Extroversion and Conscientiousness, we see that while there doesn't appear to be a big difference between males and females in younger populations, older populations have a slightly more pronounced difference with females on average being more extroverted and agreeable. Looking at the T-test results, these two populations are also significantly different, though the actual difference is slight with males having a mean of 2.69 and females having a mean of 2.76.




While the difference may be slight for sex, there does appear to be something interesting happening between countries. The Philippines regression line looks very different from all of the other countries. It seems, while younger populations of the Philippines are similar to other countries' younger populations, older Philippines populations tend to be more extroverted and conscientious. This could be another direction for further research and analysis.




```
##
##  Welch Two Sample t-test
##
## data:  males$OCEANavg and females$OCEANavg
## t = -23.867, df = 15879, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1338724 -0.1135522
## sample estimates:
## mean of x mean of y
##  3.006661  3.130373
```

Lastly, I decided to average all of the traits and see if there were any relationships among the responses. The data indicates that women appear to rate themselves higher than men when averaging all traits, and the T test indicates this is a significant difference in populations. That being said, while this average number is fun, it's not very explanatory. It does not indicate a preference towards rating all questions highly, as we corrected the data for that bias earlier. It also does not explain which traits cause men to be lower than women. Only that, on average, men tend to be more introverted, less agreeable, less conscientious, less neurotic, and less open to experiences than women as a whole.

## Topics From Class

(a) Git
(b) RMarkdown
(c) Statistical concepts such as normal distributions, mean, standard deviations, percentiles, and areas under the curve
(d) Regression lines to find trends in trait expression.

(e) T tests to determine if two populations are significantly different.

## Conclusion

In conclusion, we discovered that men and women have a few significant differences between trait expression. Women tend to self report as being more extroverted, agreeable, and neurotic than men. We also discovered that English native speakers self report as being slightly more open to experience and less neurotic than their non-native counterparts. We determined that, while most countries most of the time have little difference in personality expression, the Philippines is an outlier with some interesting trends in trait expression and age. Lastly, it appears that across the board there are more responses and those responses are more varried with younger populations. While all of these findings are interesting, it's important to note that this is a self-reported online test. It does not reflect a true random sampling, and the responses could be biased based on inaccurate self conceptions of respondents or false responses by participants. The participants were informed that the results of this test would be used for scientific purposes.

Once again, this dataset was not created by myself, and can be found at Kaggle