SEIS631 Final Project

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.

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.

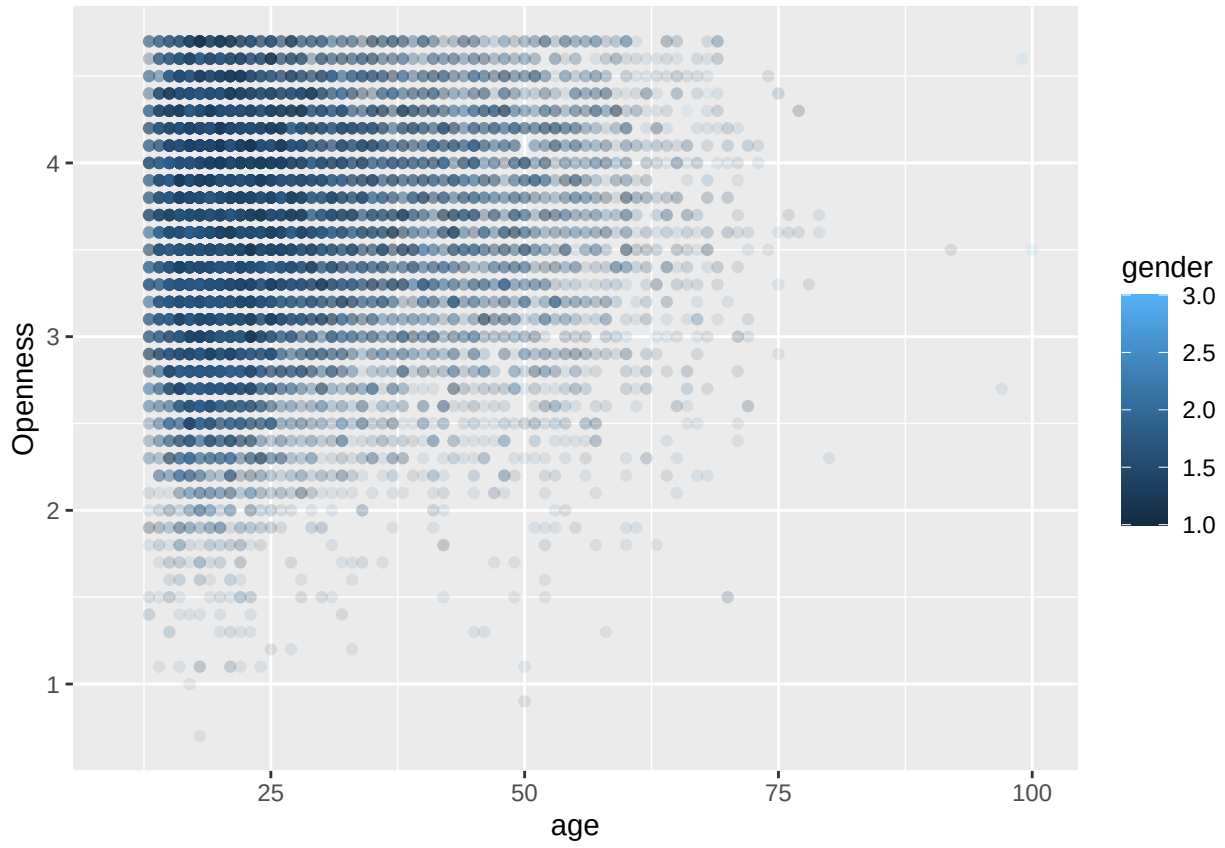
Body

Cleaning the data.

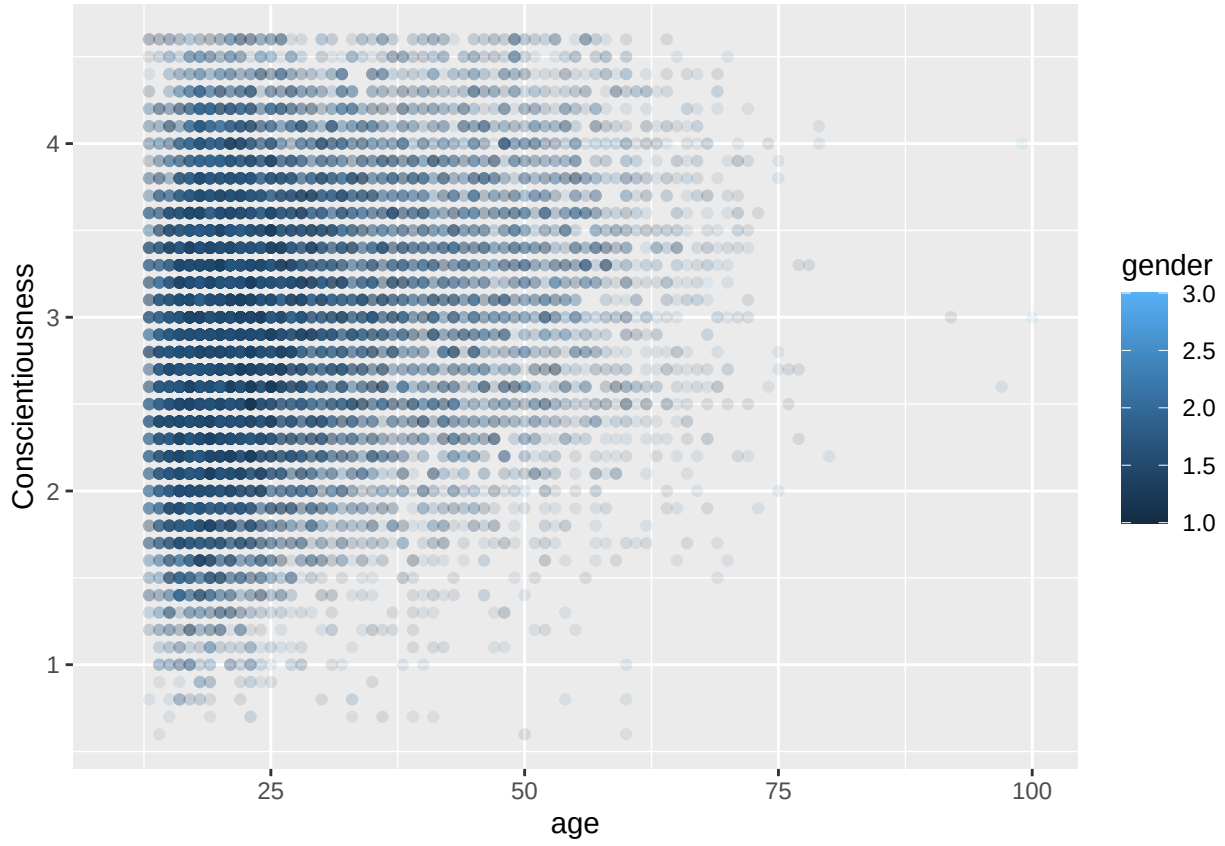
The dataset was quite large and had demographic information and the actual responses to the individual questions. While this is ideal, it means the dataset must be cleaned to be useful. For example, to account for the tendency to give high ratings over low ratings regardless of the question's content, 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.

```
## Warning: Removed 83 rows containing missing values (geom_point).
```

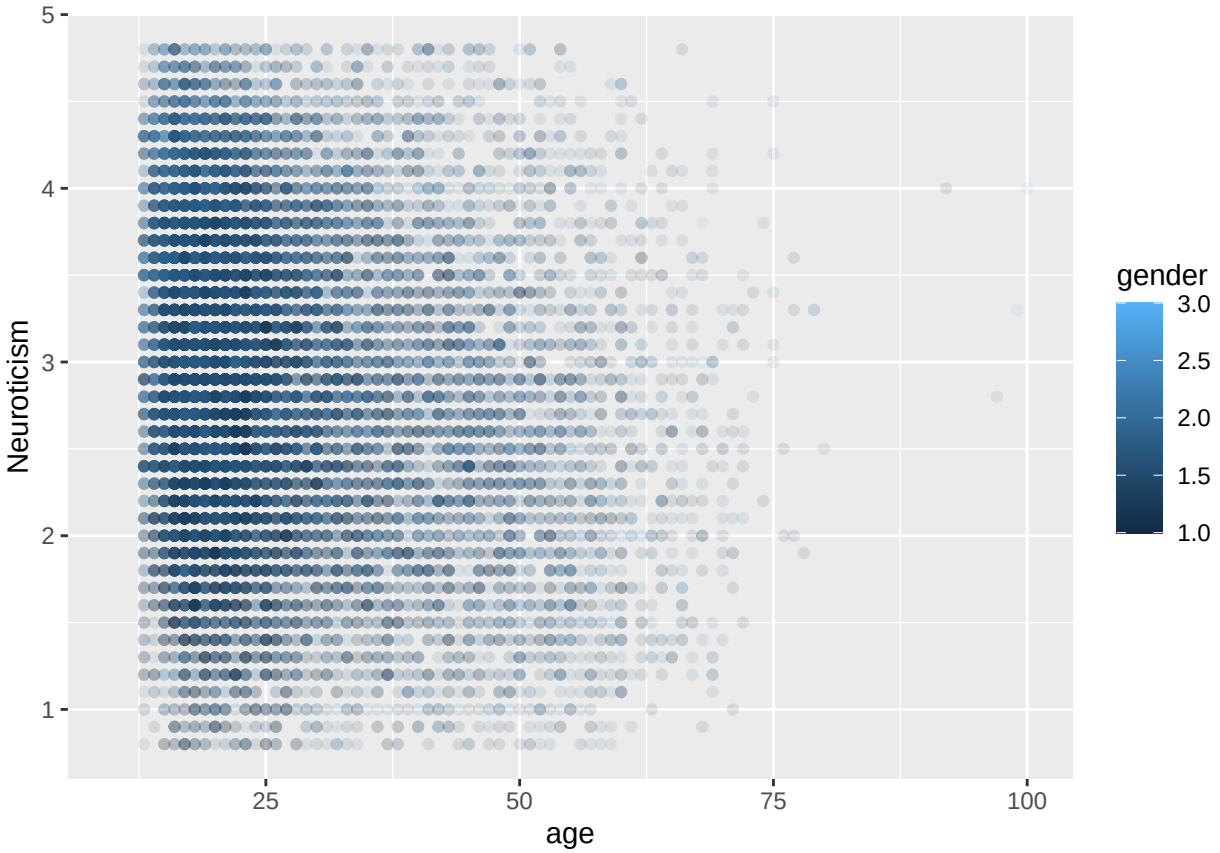

```
## Warning: Removed 83 rows containing missing values (geom_point).
```



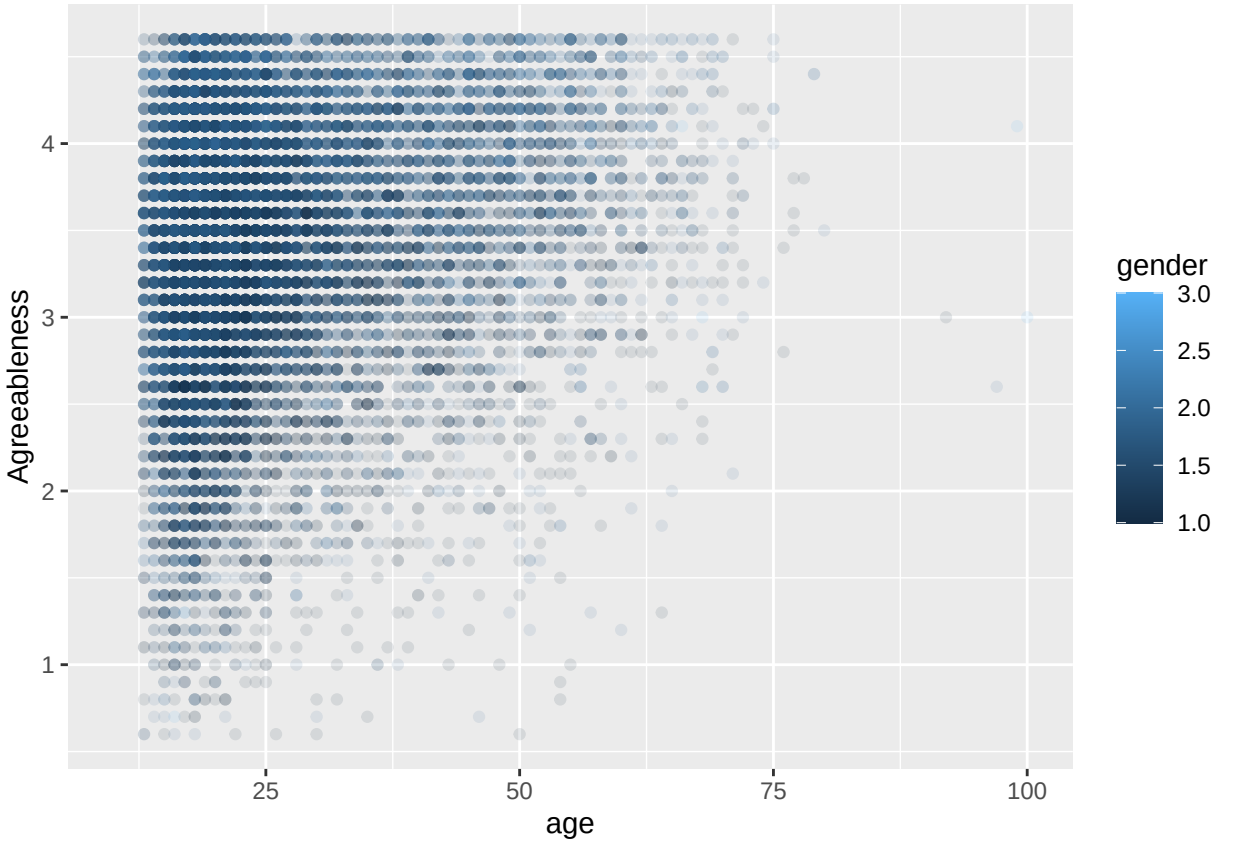

```
## Warning: Removed 83 rows containing missing values (geom_point).
```




```
## Warning: Removed 83 rows containing missing values (geom_point).
```

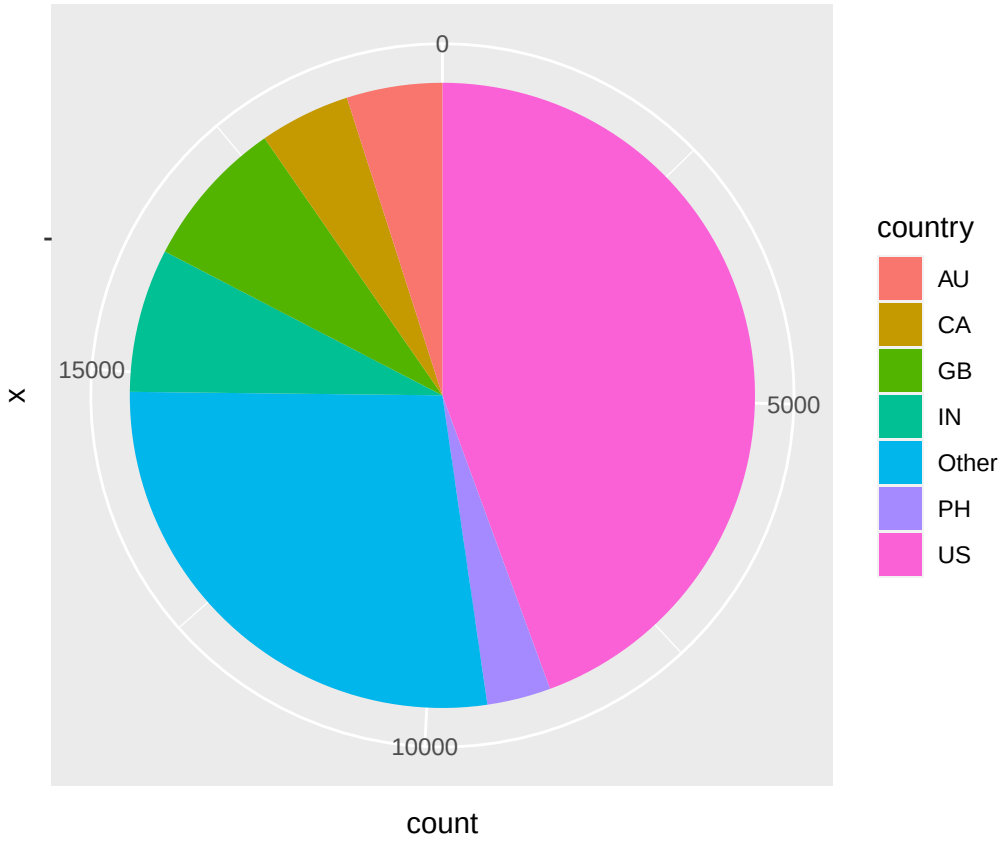


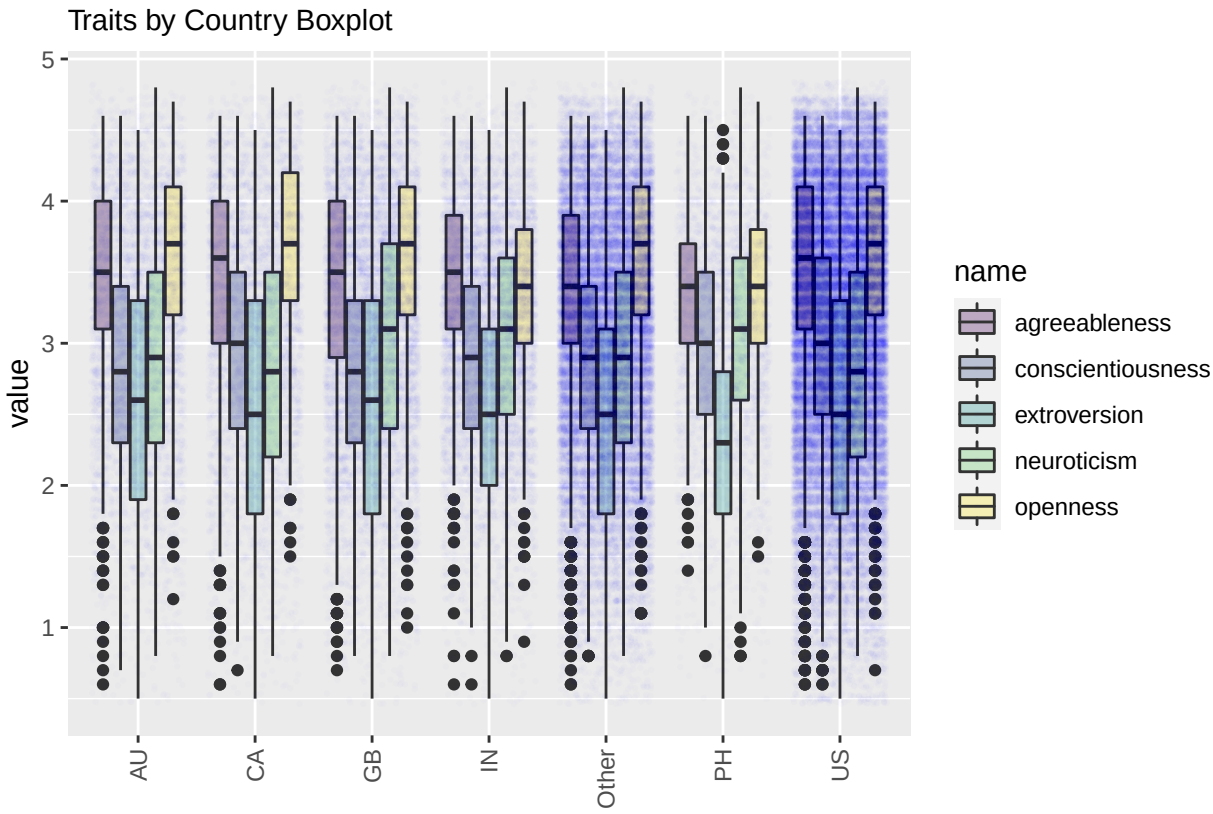

```
## Warning: Removed 83 rows containing missing values (geom_point).
```

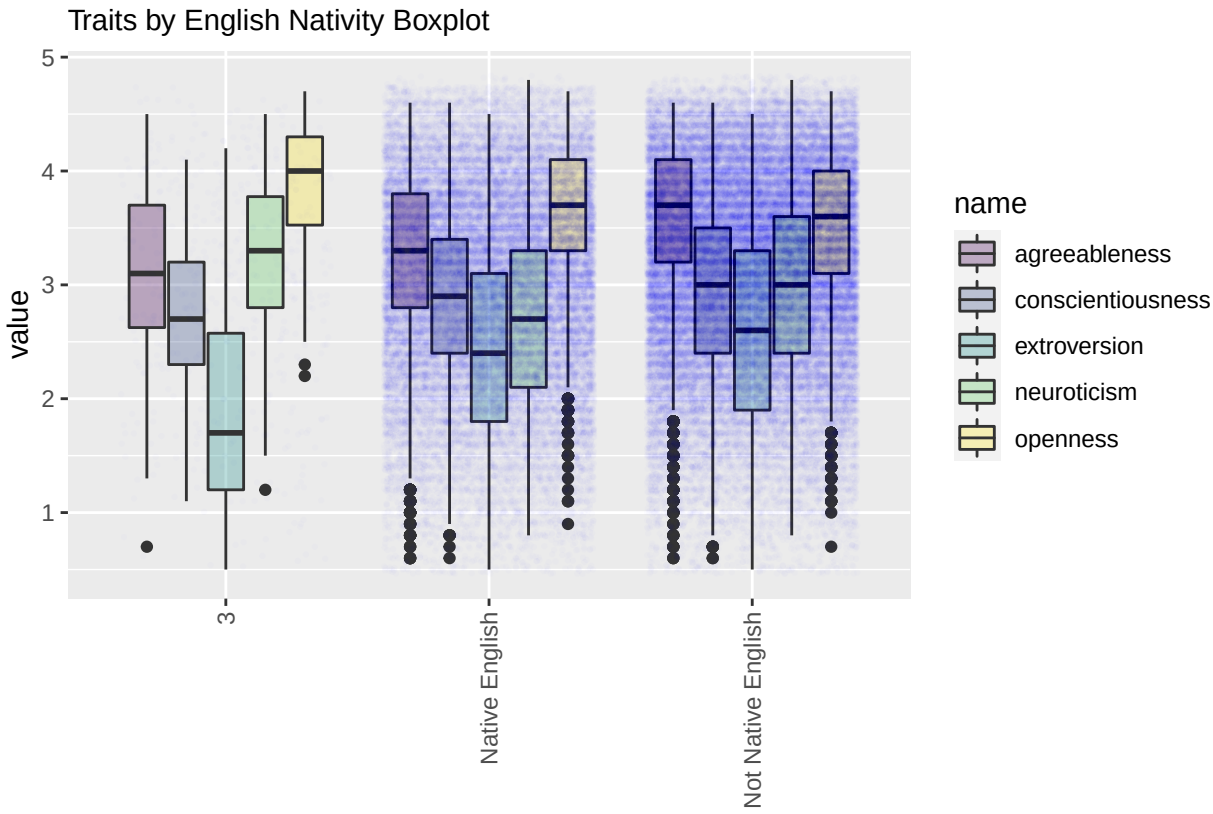


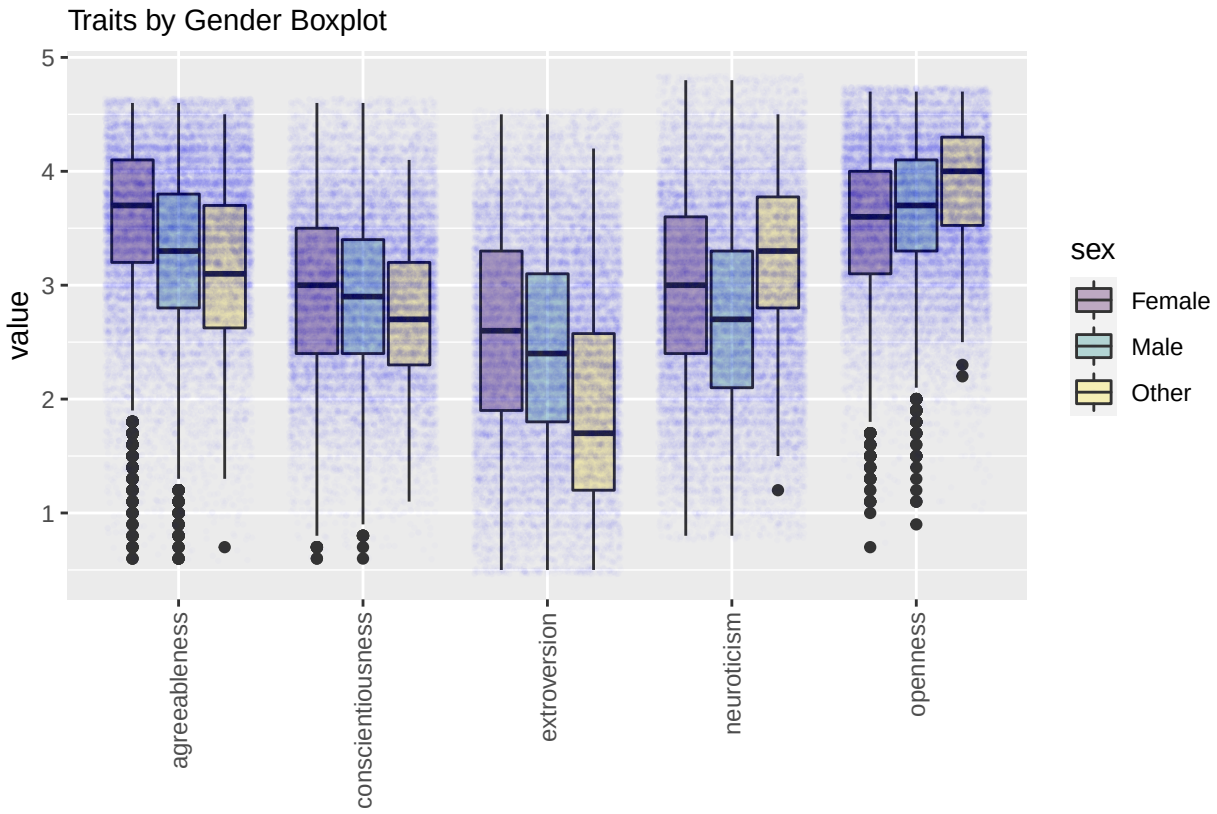
If we take the averages of all of the question's scores for each trait for respondents in the US, we see that most respondents were age 25 or younger. We also see trends, such as Agreeableness and Openness tend to be higher than 2. Conscientiousness and Neuroticism seem to avoid the valences of 1 and 5.Extroversion, on the other hand, appears to be across the board.looking at the above graph of age versus extroversion, it seems there is little relation between age and being introverted or extroverted with the exception of after age 60 it appears like less people are strongly introverted.

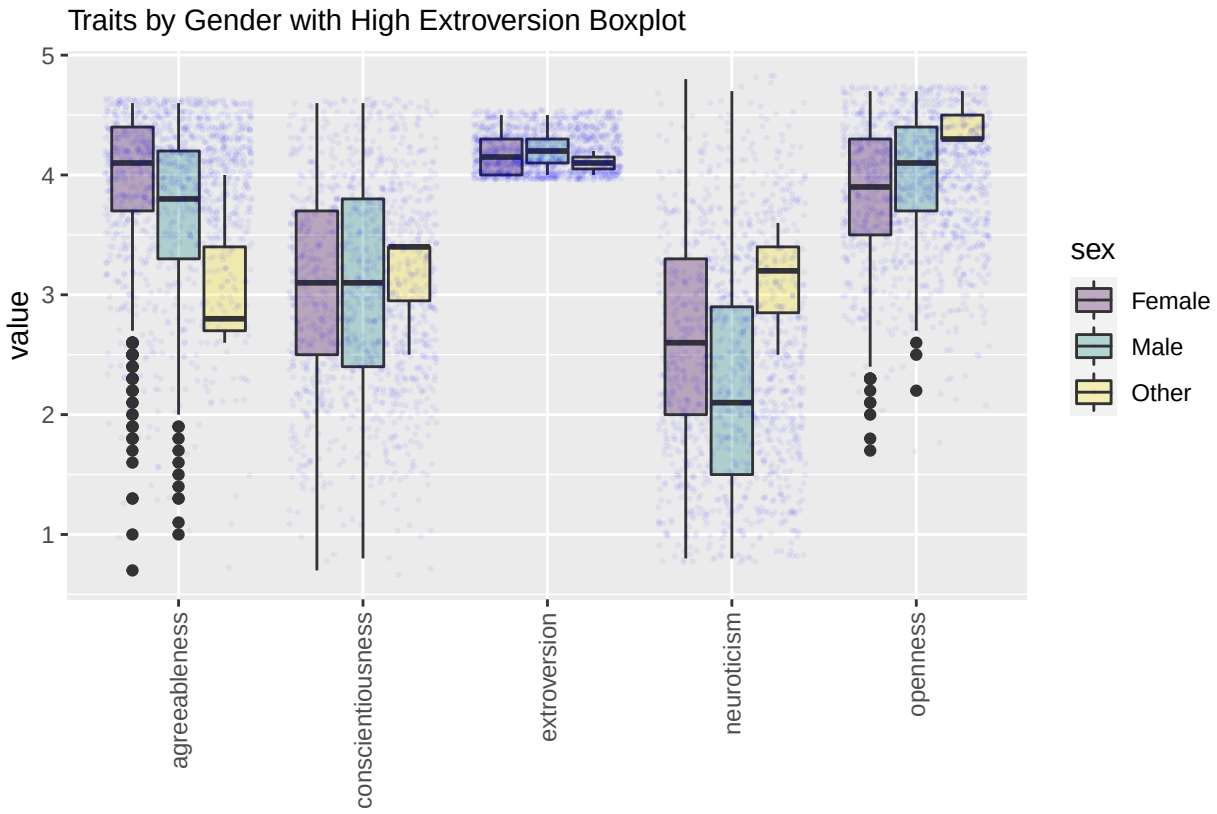
As seen by this R code, the top countries with observations in this dataset are English speaking countries. I will be looking at these countries next and similarities and differences within these countries to see if country, age, and sex result in significant differences.

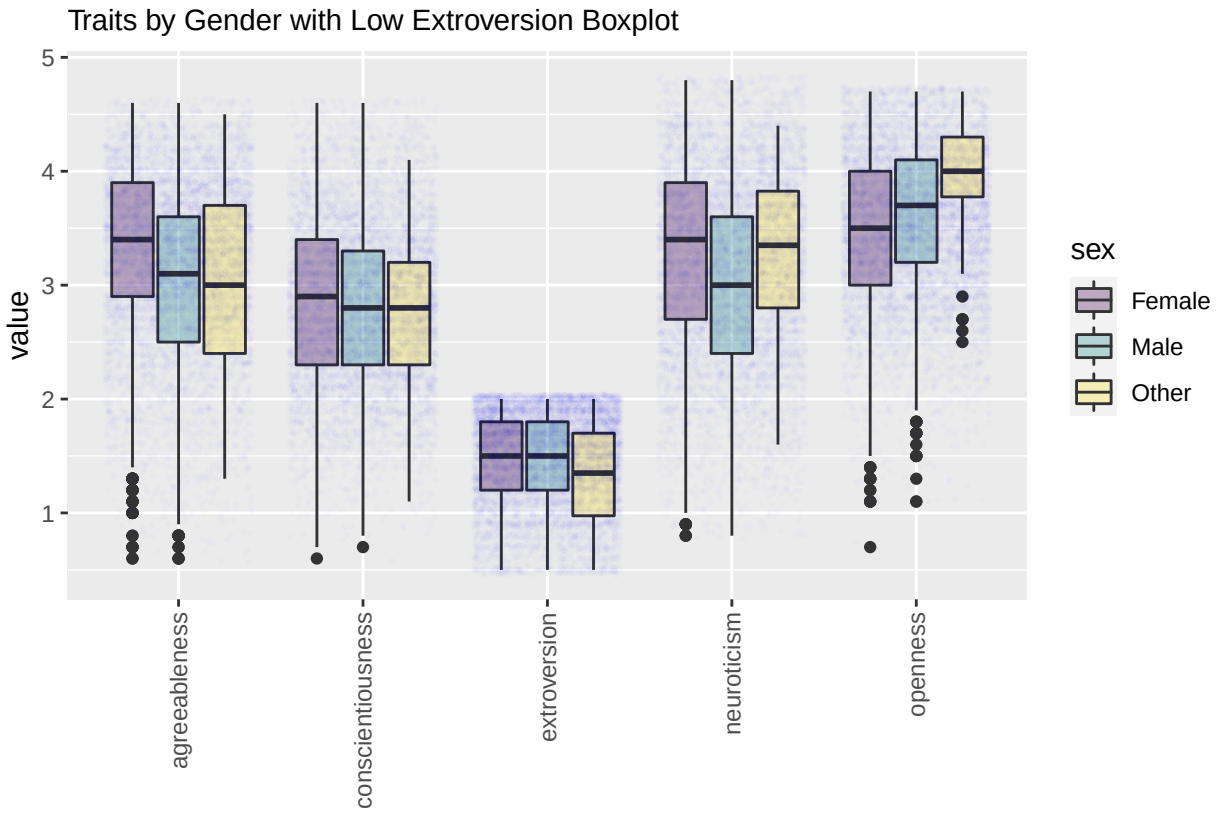

0
5000
10000
15000
x
count
country
AU
CA
GB
IN
Other
PH
US

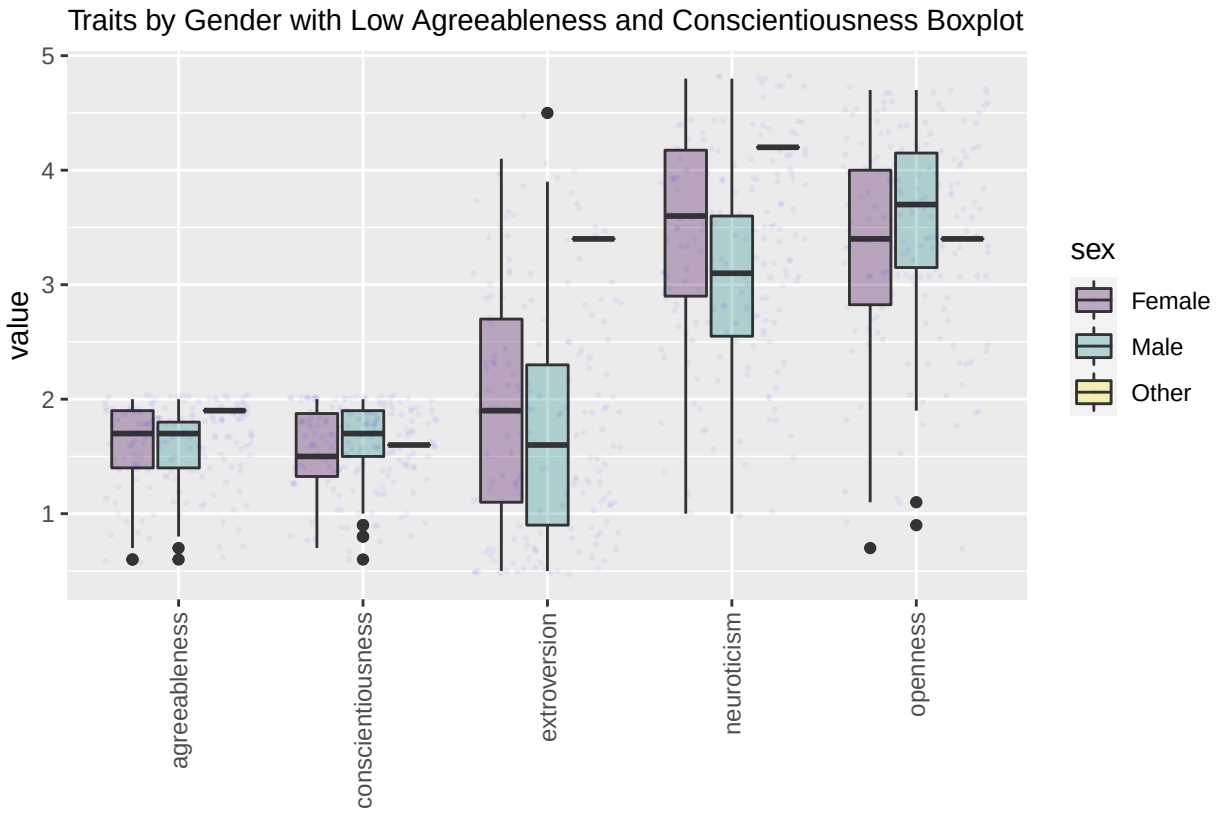

Traits by Country Boxplot
value
5
4
3
2
1
AU
CA
GB
IN
Other
PH
US
name
agreeableness
conscientiousness
extroversion
neuroticism
openness

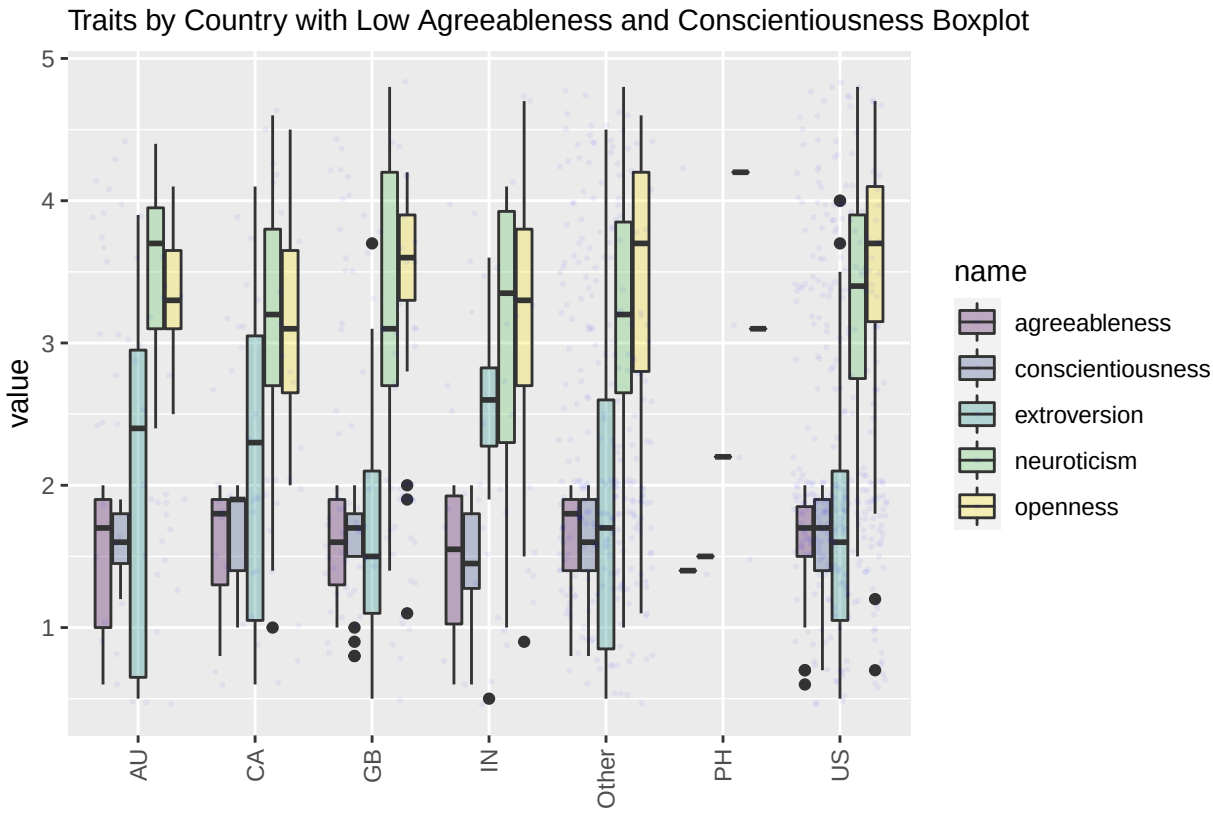

Traits by English Nativity Boxplot
value
5
4
3
2
1
3
Native English
Not Native English
name
agreeableness
conscientiousness
extroversion
neuroticism
openness

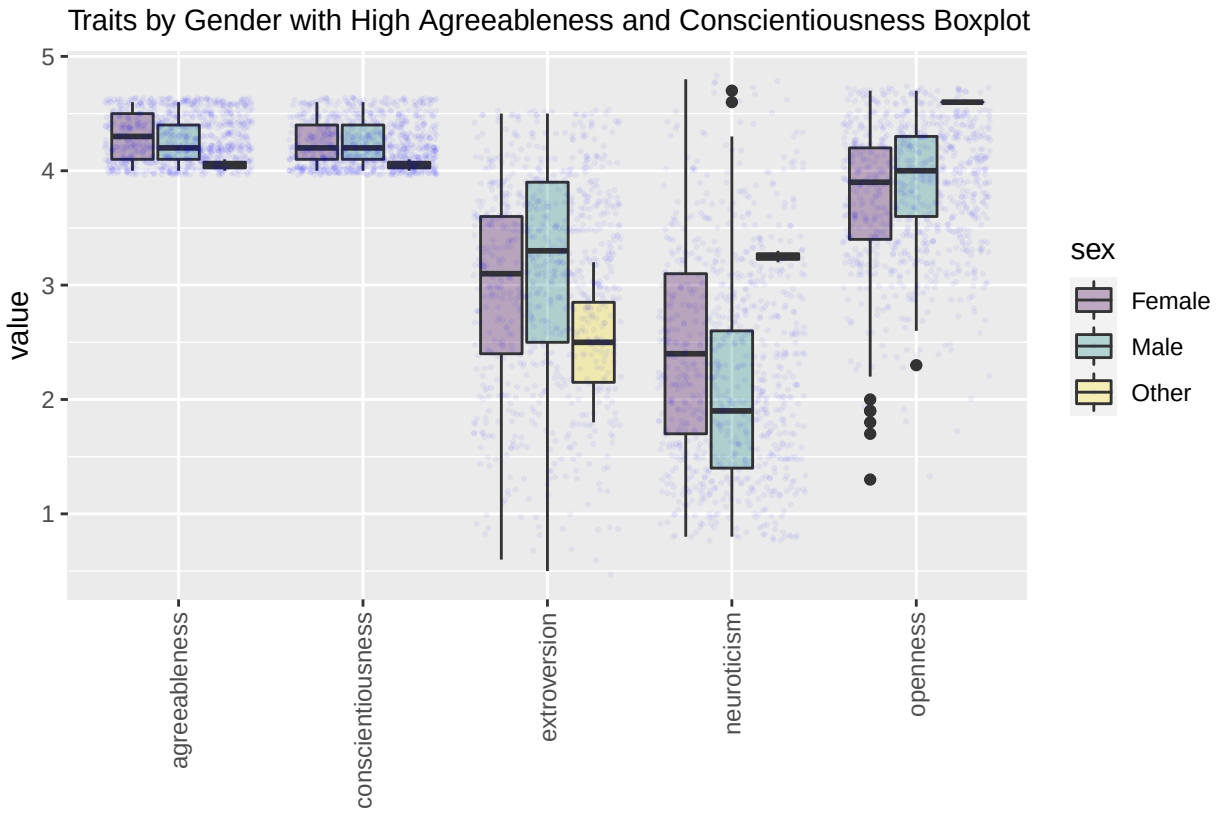

Traits by Gender Boxplot
value
5
4
3
2
1
agreeableness
conscientiousness
extroversion
neuroticism
openness
sex
Female
Male
Other

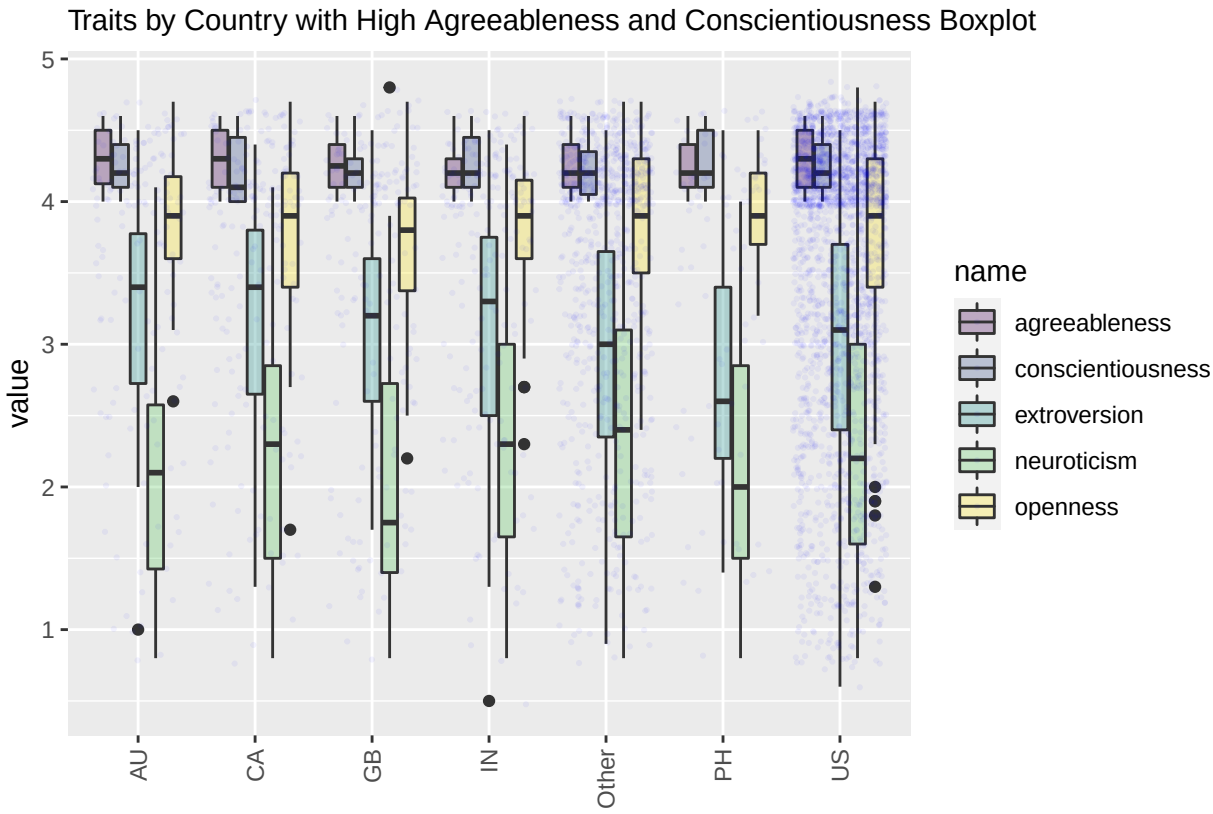

Traits by Gender with High Extroversion Boxplot
value
5
4
3
2
1
agreeableness
conscientiousness
extroversion
neuroticism
openness
sex
Female
Male
Other

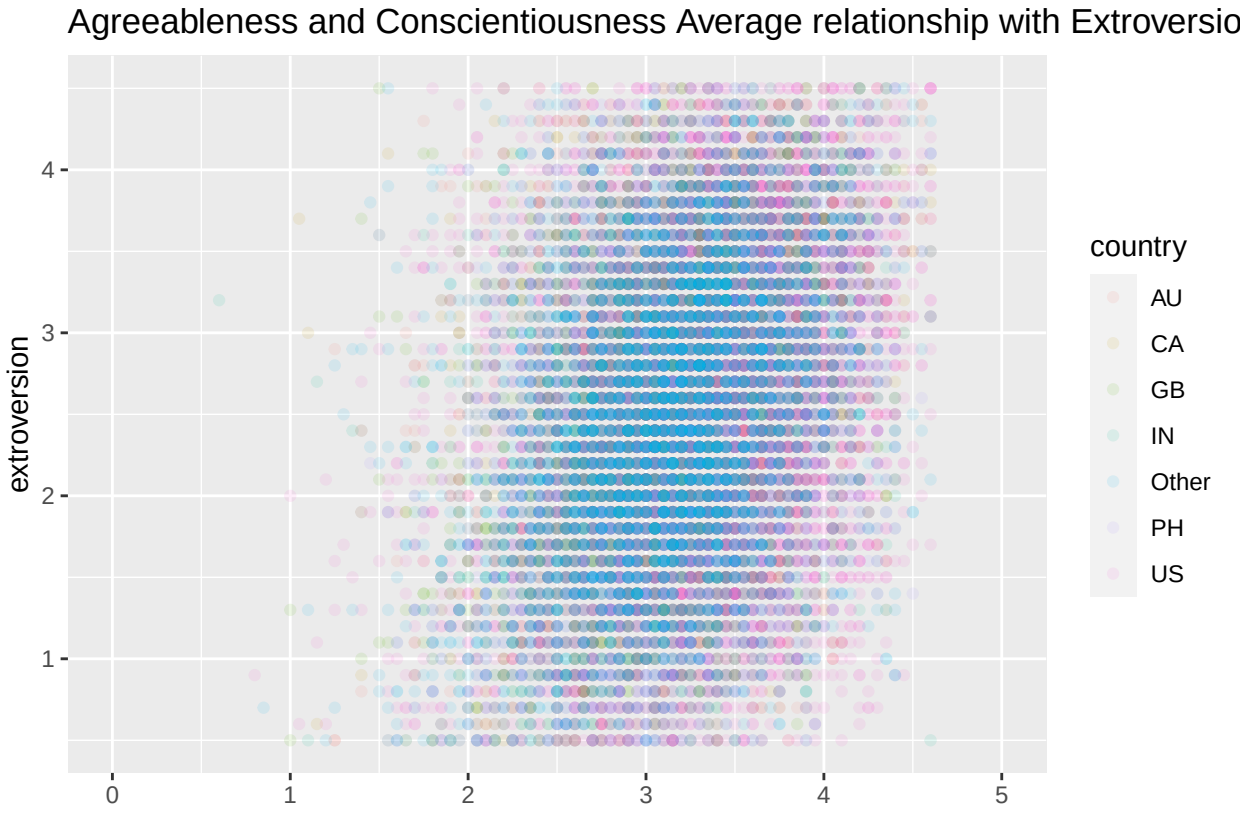

Traits by Gender with Low Extroversion Boxplot
value
5
4
3
2
1
agreeableness
conscientiousness
extroversion
neuroticism
openness
sex
Female
Male
Other

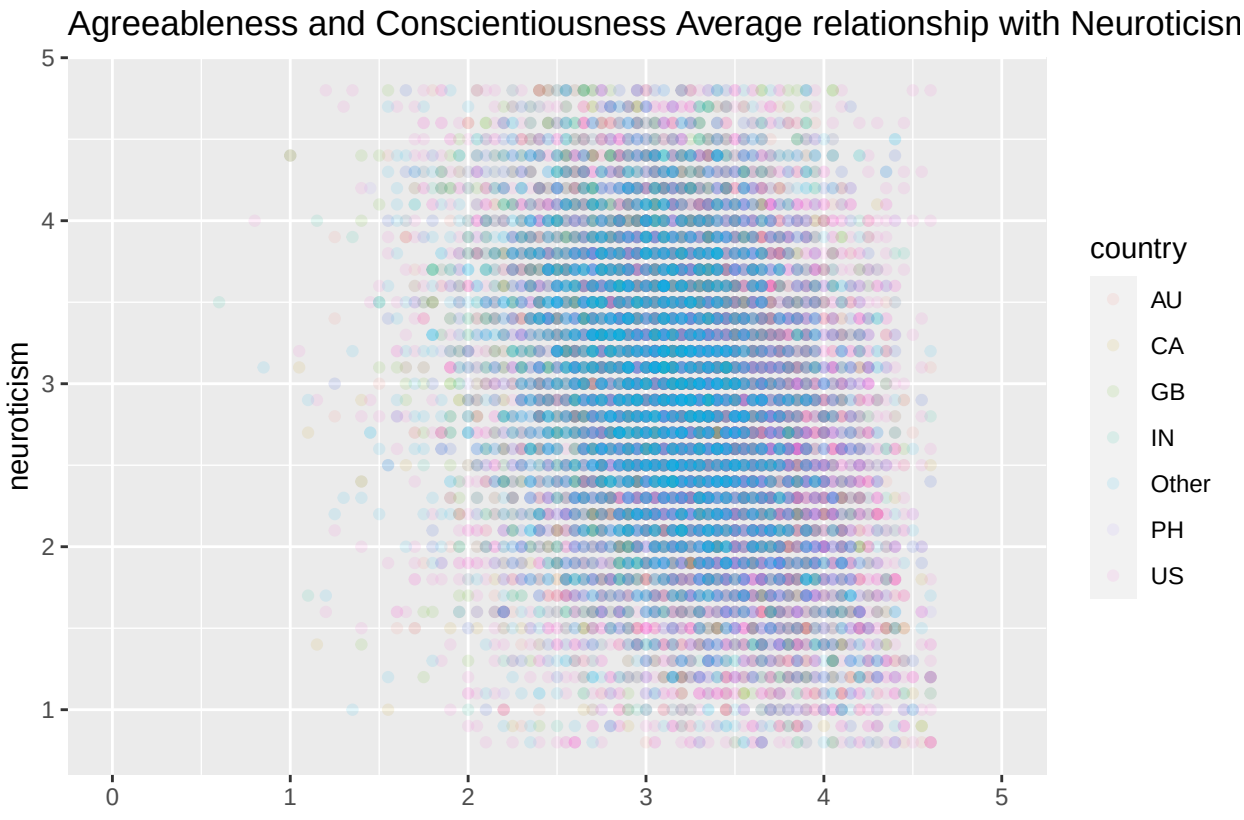

Traits by Gender with Low Agreeableness and Conscientiousness Boxplot
value
5
4
3
2
1
agreeableness
conscientiousness
extroversion
neuroticism
openness
sex
Female
Male
Other

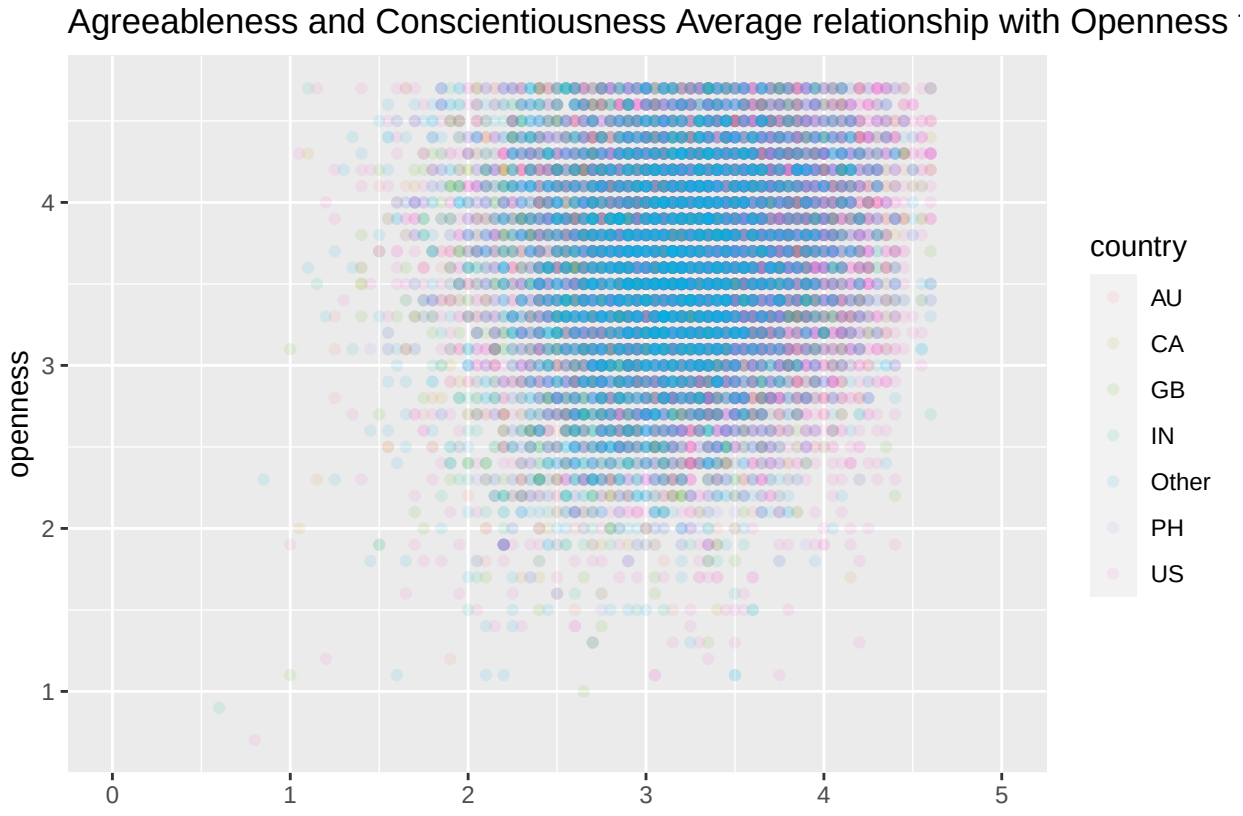

Traits by Country with Low Agreeableness and Conscientiousness Boxplot
value
5
4
3
2
1
AU
CA
GB
IN
Other
PH
US
name
agreeableness
conscientiousness
extroversion
neuroticism
openness

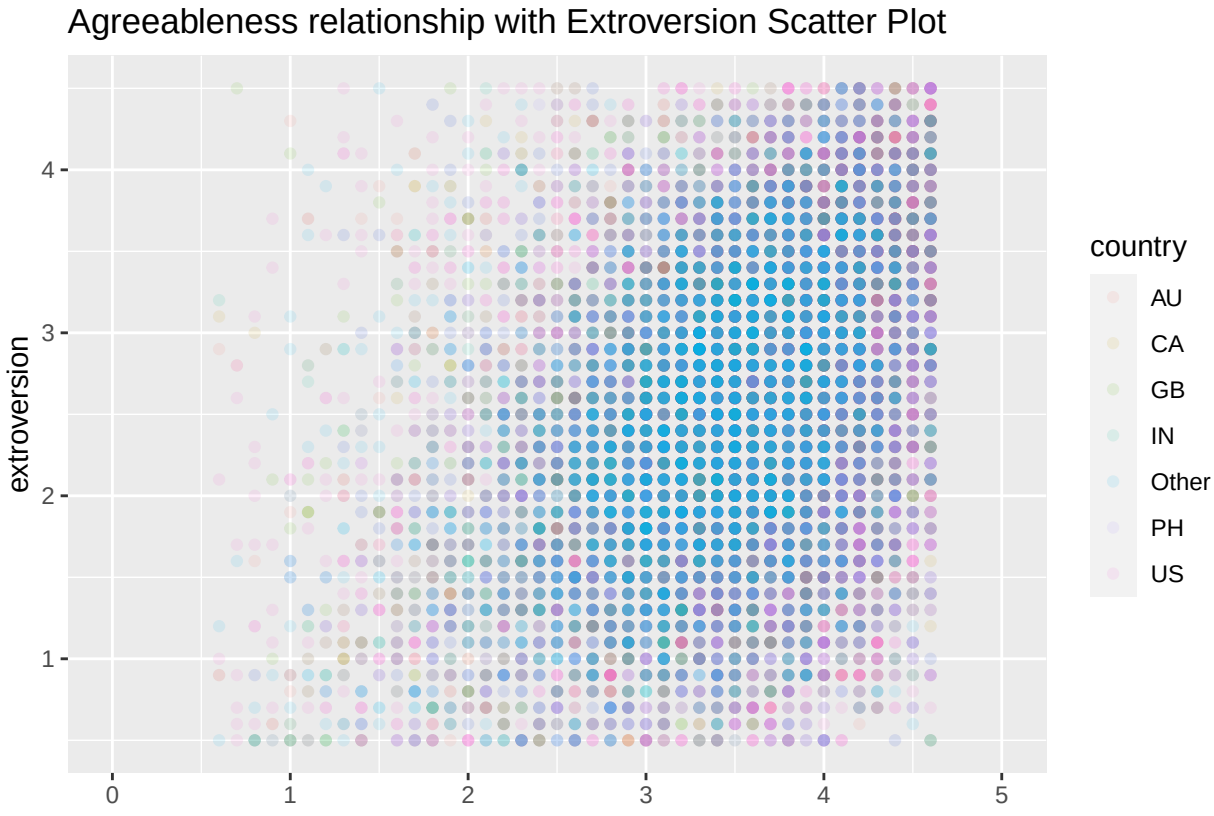

Traits by Gender with High Agreeableness and Conscientiousness Boxplot
value
5
4
3
2
1
agreeableness
conscientiousness
extroversion
neuroticism
openness
sex
Female
Male
Other

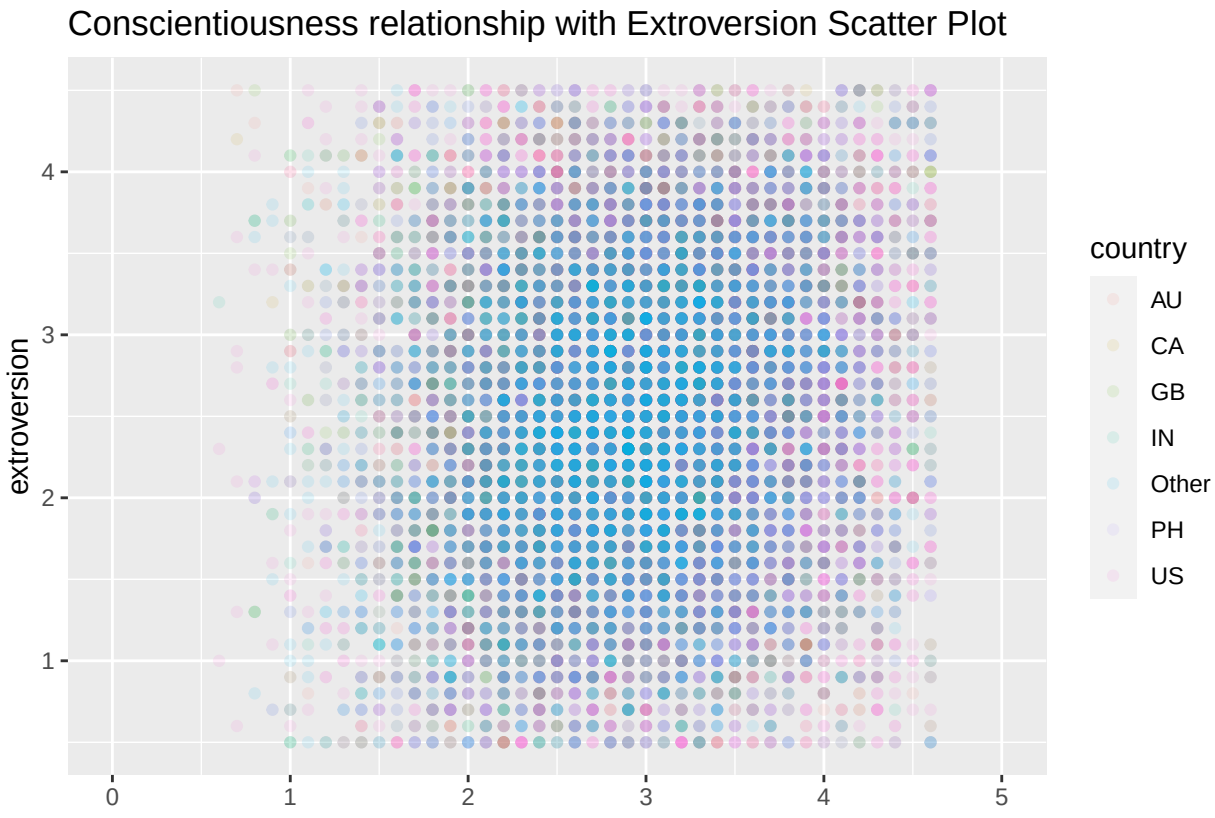

Traits by Country with High Agreeableness and Conscientiousness Boxplot
value
5
4
3
2
1
AU
CA
GB
IN
Other
PH
US
name
agreeableness
conscientiousness
extroversion
neuroticism
openness

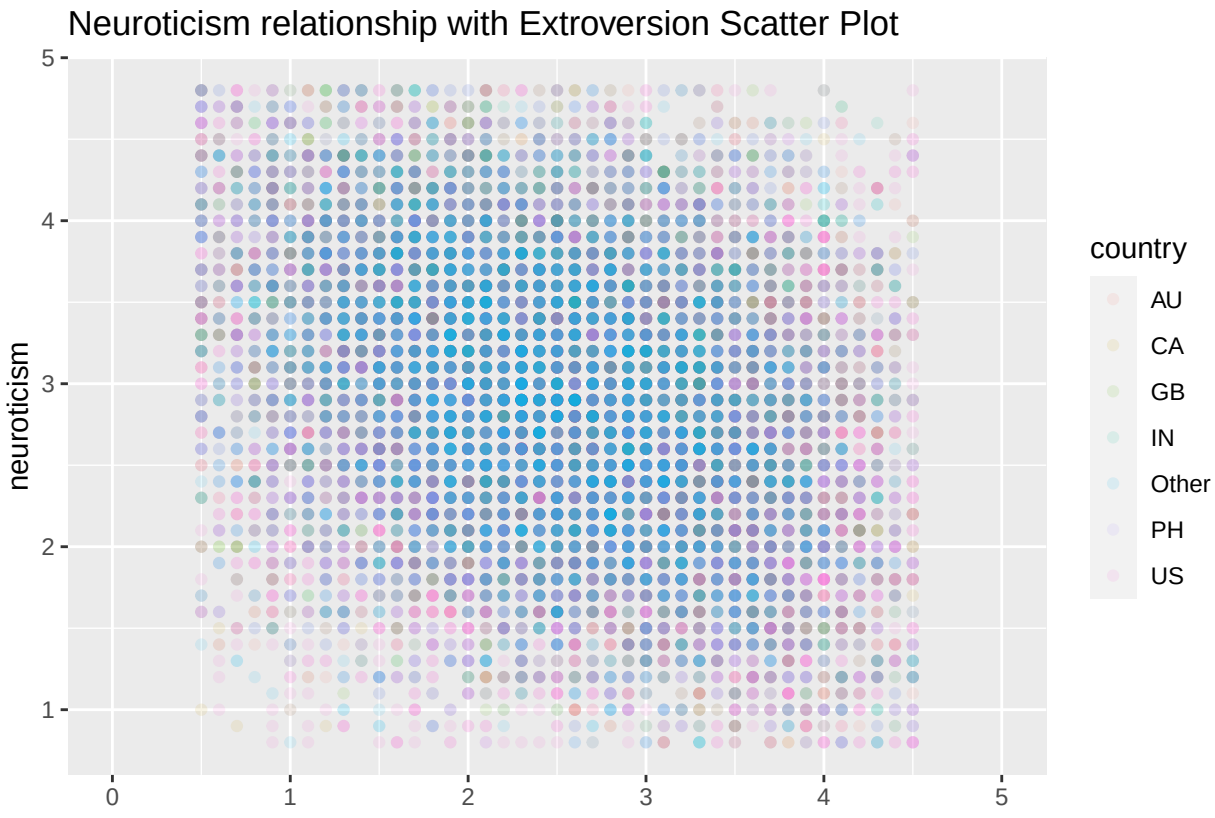

Agreeableness and Conscientiousness Average relationship with Extroversio
extroversion
4
3
2
1
0
1
2
3
4
5
country
AU
CA
GB
IN
Other
PH
US

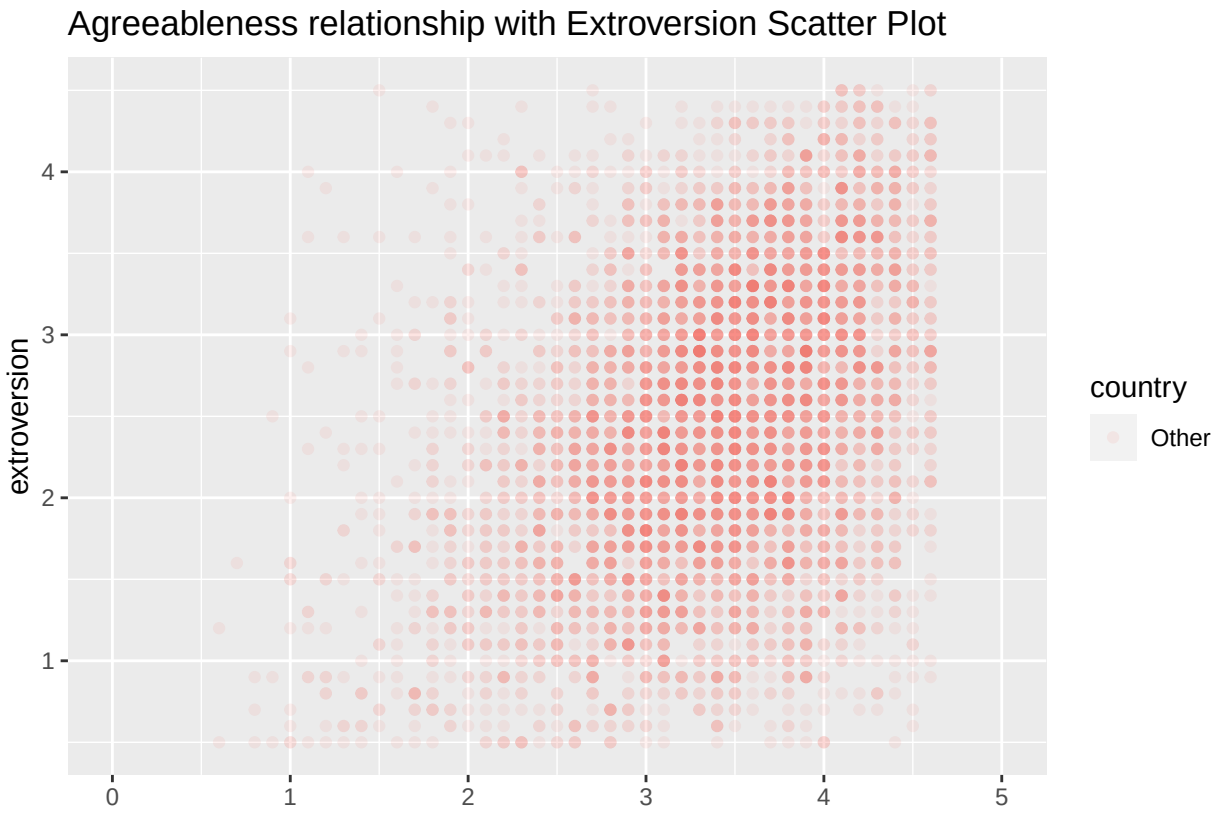

Agreeableness and Conscientiousness Average relationship with Neuroticism
neuroticism
5
4
3
2
1
0
1
2
3
4
5
country
AU
CA
GB
IN
Other
PH
US

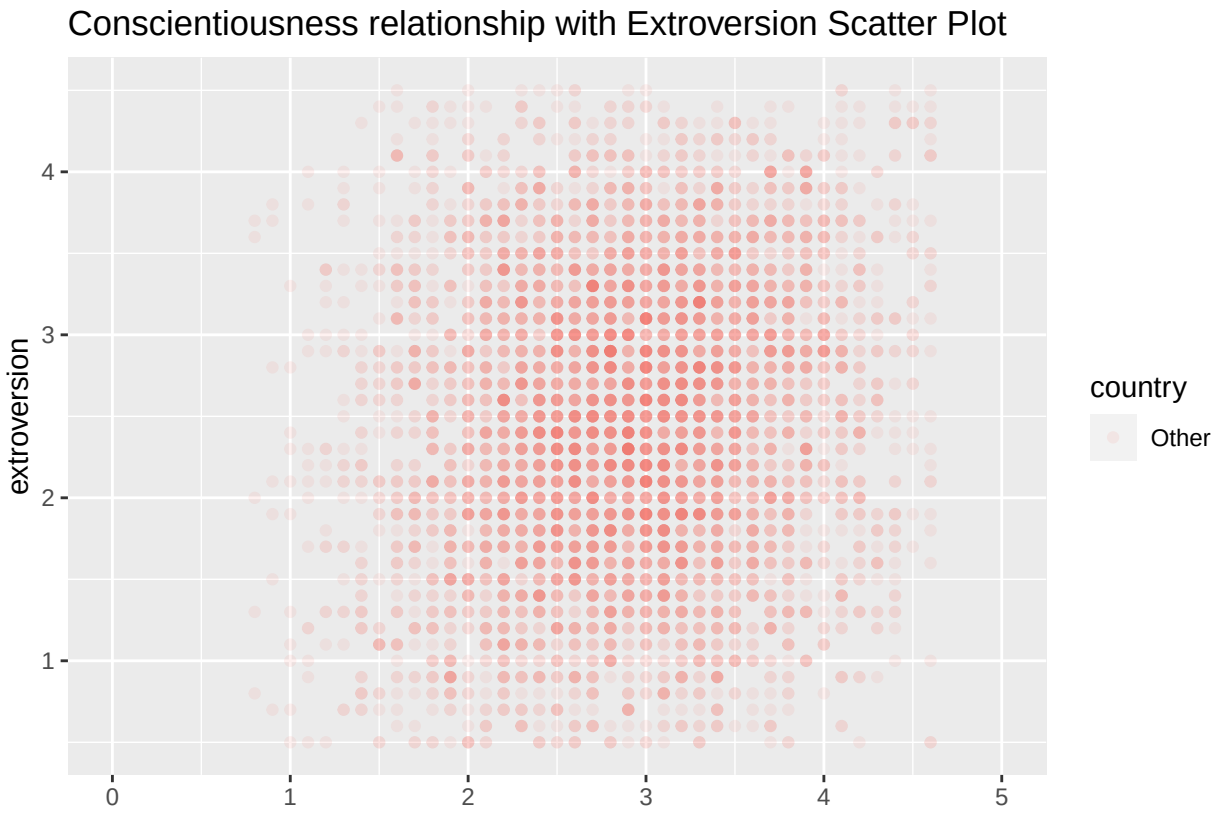

Agreeableness and Conscientiousness Average relationship with Openness
openness
4
3
2
1
0
1
2
3
4
5
country
AU
CA
GB
IN
Other
PH
US

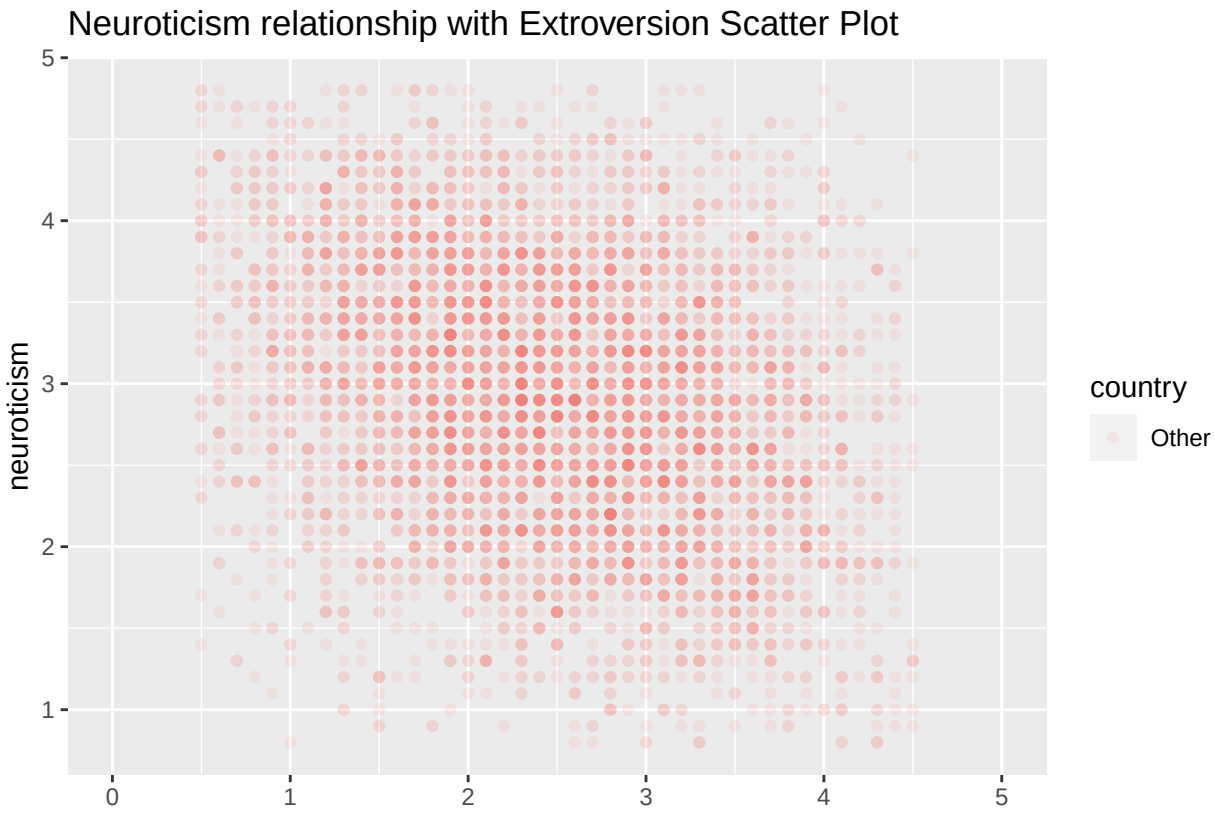

Agreeableness relationship with Extroversion Scatter Plot
extroversion
4
3
2
1
0
1
2
3
4
5
country
AU
CA
GB
IN
Other
PH
US


Conscientiousness relationship with Extroversion Scatter Plot
extroversion
4
3
2
1
0
1
2
3
4
5
country
AU
CA
GB
IN
Other
PH
US


Neuroticism relationship with Extroversion Scatter Plot
neuroticism
5
4
3
2
1
0
1
2
3
4
5
country
AU
CA
GB
IN
Other
PH
US

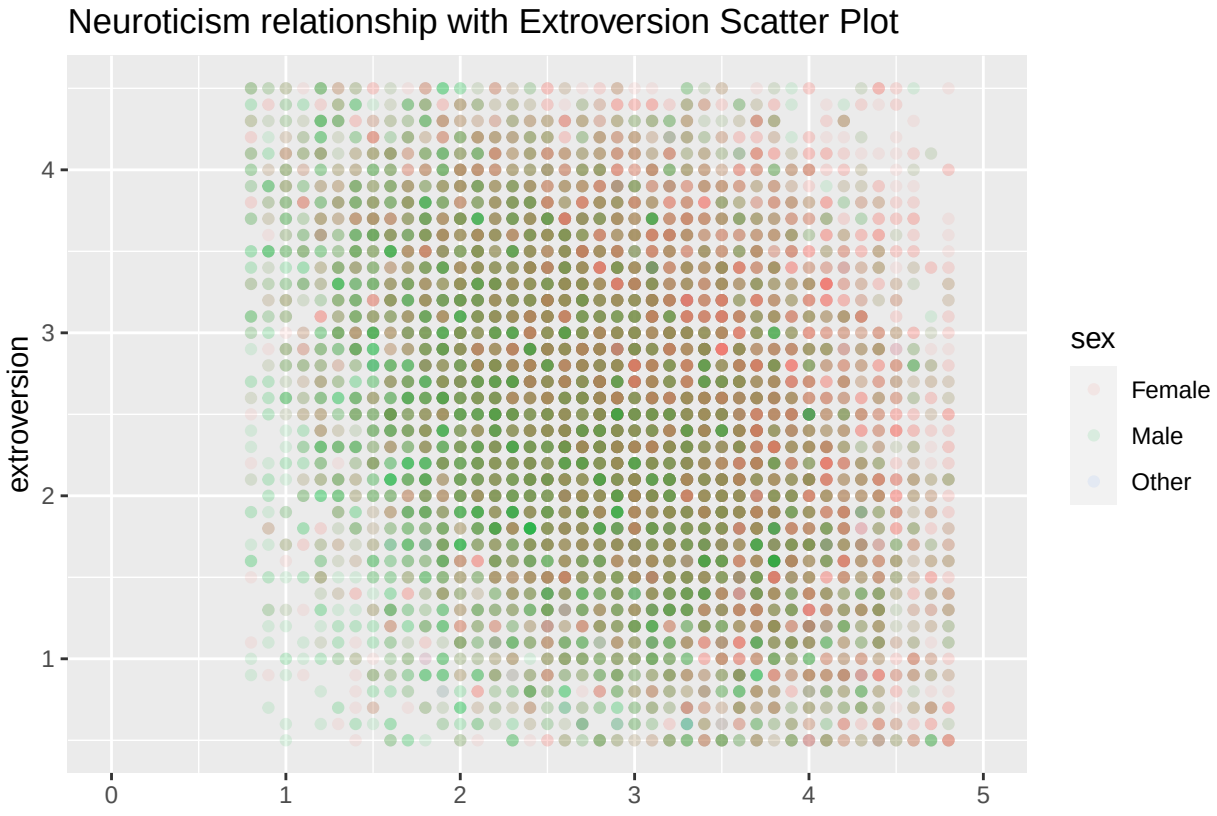

Agreeableness relationship with Extroversion Scatter Plot
extroversion
4
3
2
1
0
1
2
3
4
5
country
Other


Conscientiousness relationship with Extroversion Scatter Plot
extroversion
4
3
2
1
0
1
2
3
4
5
country
Other

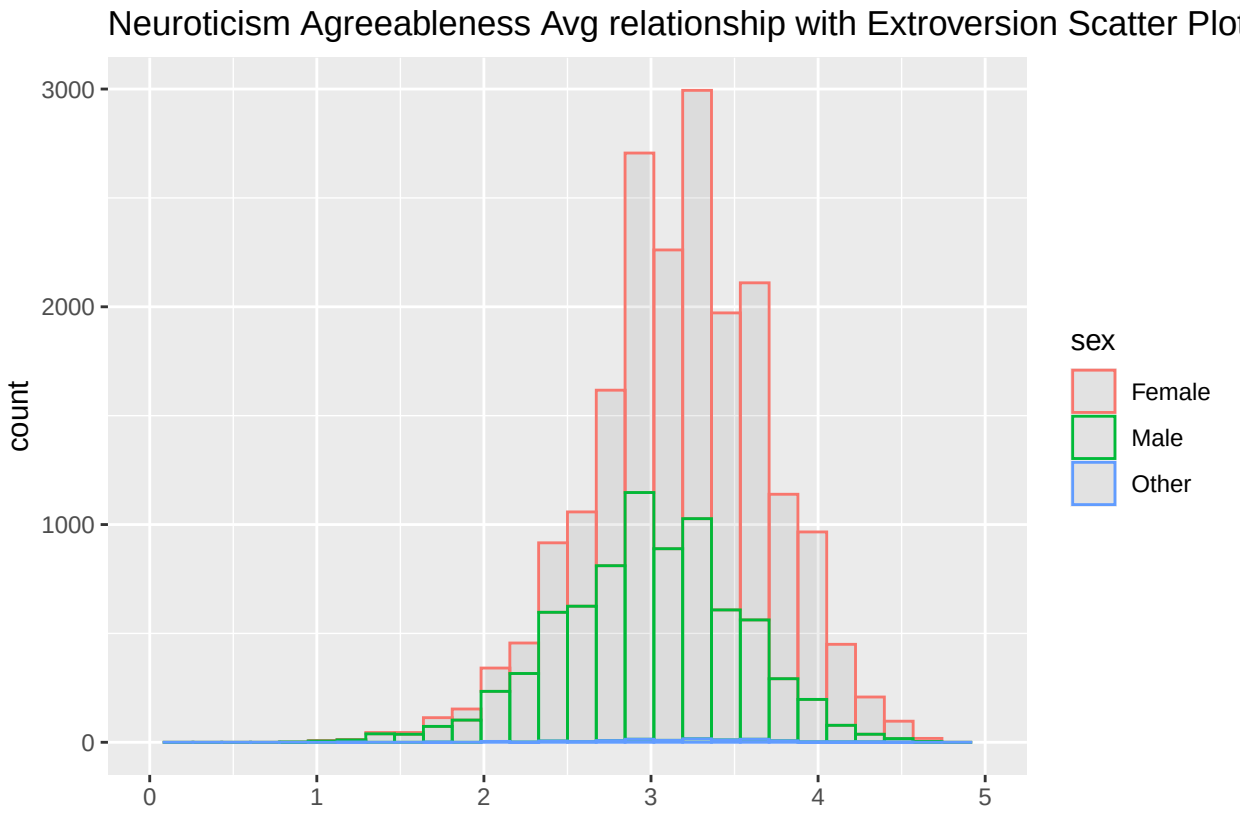

Neuroticism relationship with Extroversion Scatter Plot
neuroticism
5
4
3
2
1
0
1
2
3
4
5
country
Other

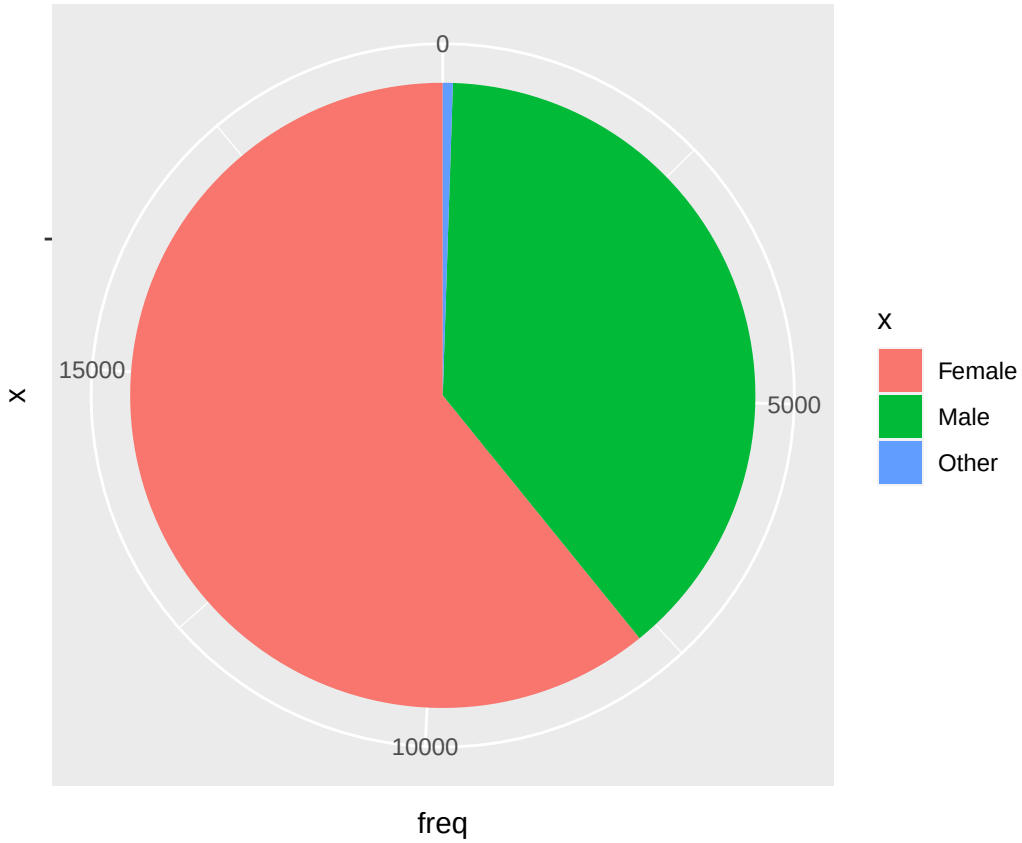

Agreeableness relationship with Extroversion Scatter Plot
extroversion
4
3
2
1
0
1
2
3
4
5
sex
Female
Male
Other


Conscientiousness relationship with Extroversion Scatter Plot
extroversion
4
3
2
1
0
1
2
3
4
5
sex
Female
Male
Other

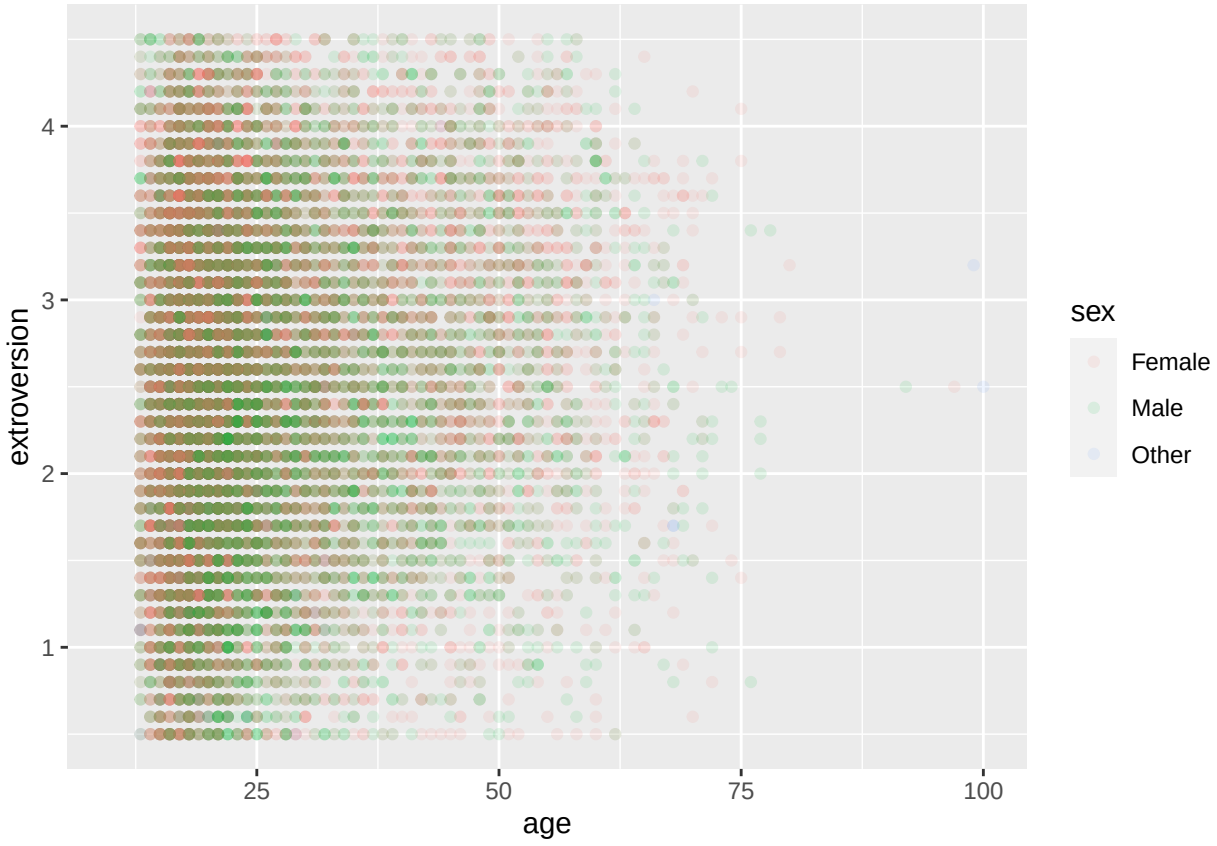

Neuroticism relationship with Extroversion Scatter Plot
extroversion
4
3
2
1
0
1
2
3
4
5
sex
Female
Male
Other


```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 6 rows containing missing values (geom_bar).
```

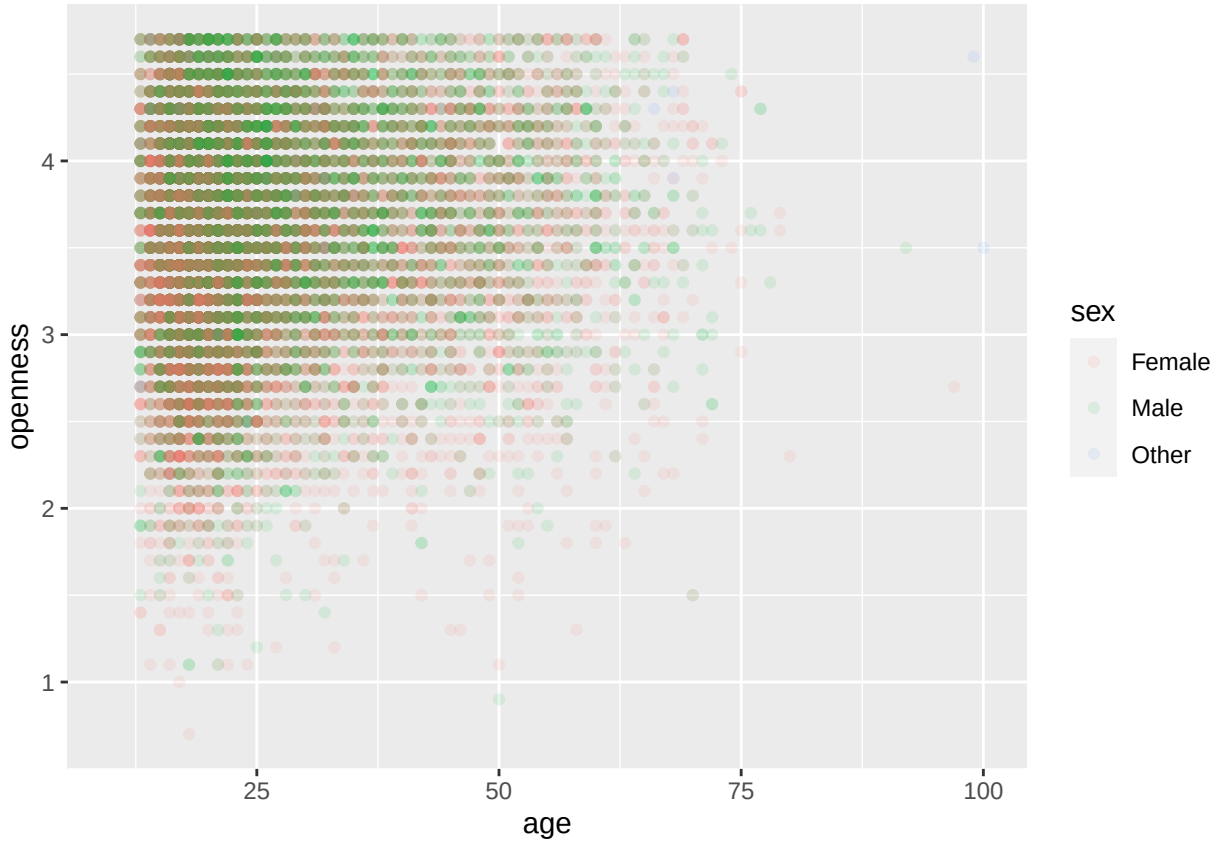

Neuroticism Agreeableness Avg relationship with Extroversion Scatter Plo
count
3000
2000
1000
0
0
1
2
3
4
5
sex
Female
Male
Other



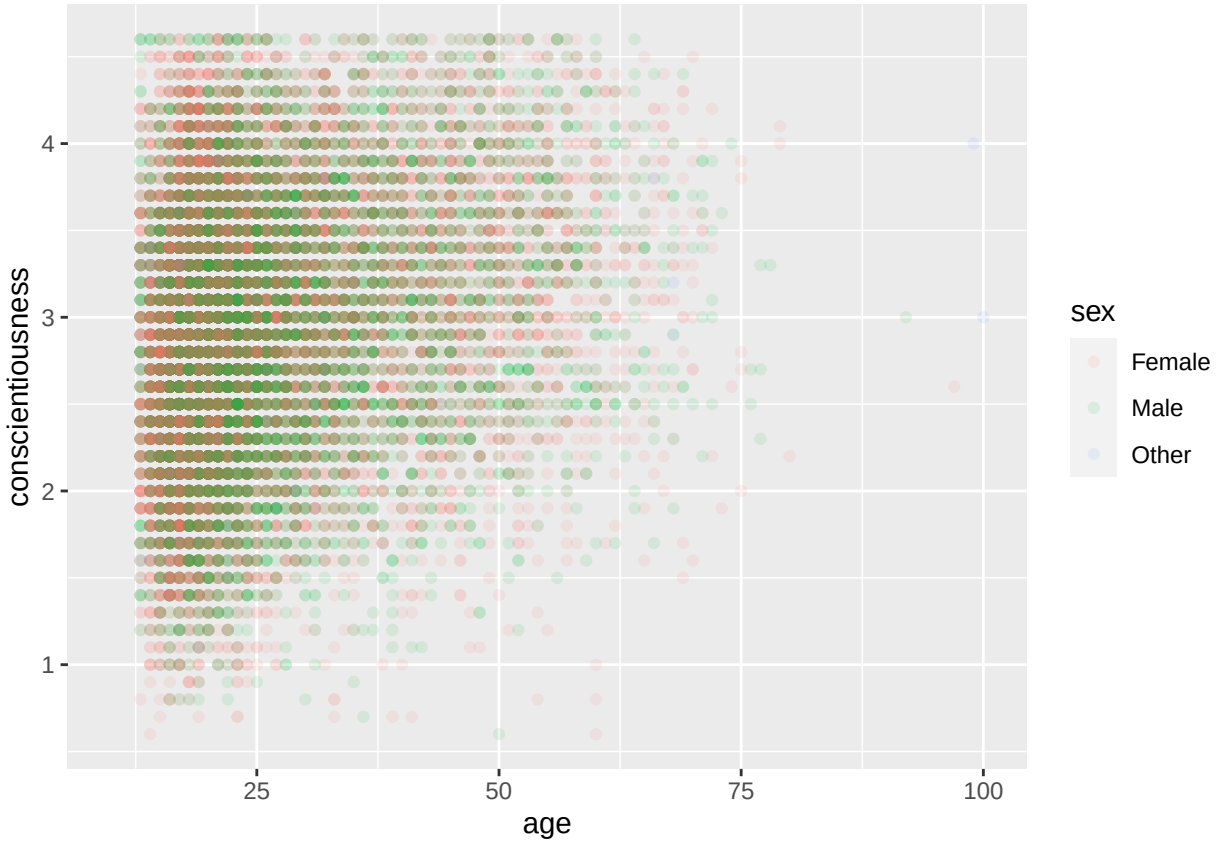

```
## Warning: Removed 83 rows containing missing values (geom_point).
```



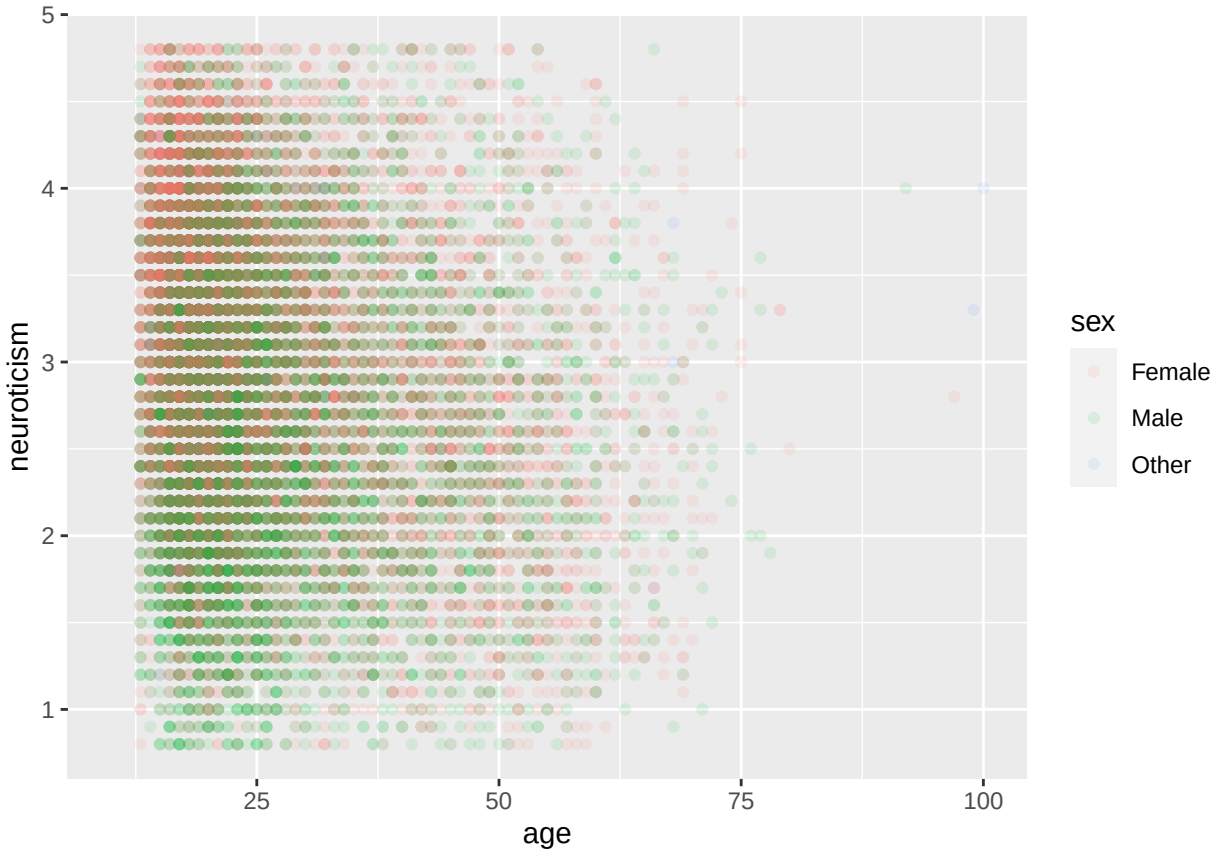

```
## Warning: Removed 83 rows containing missing values (geom_point).
```



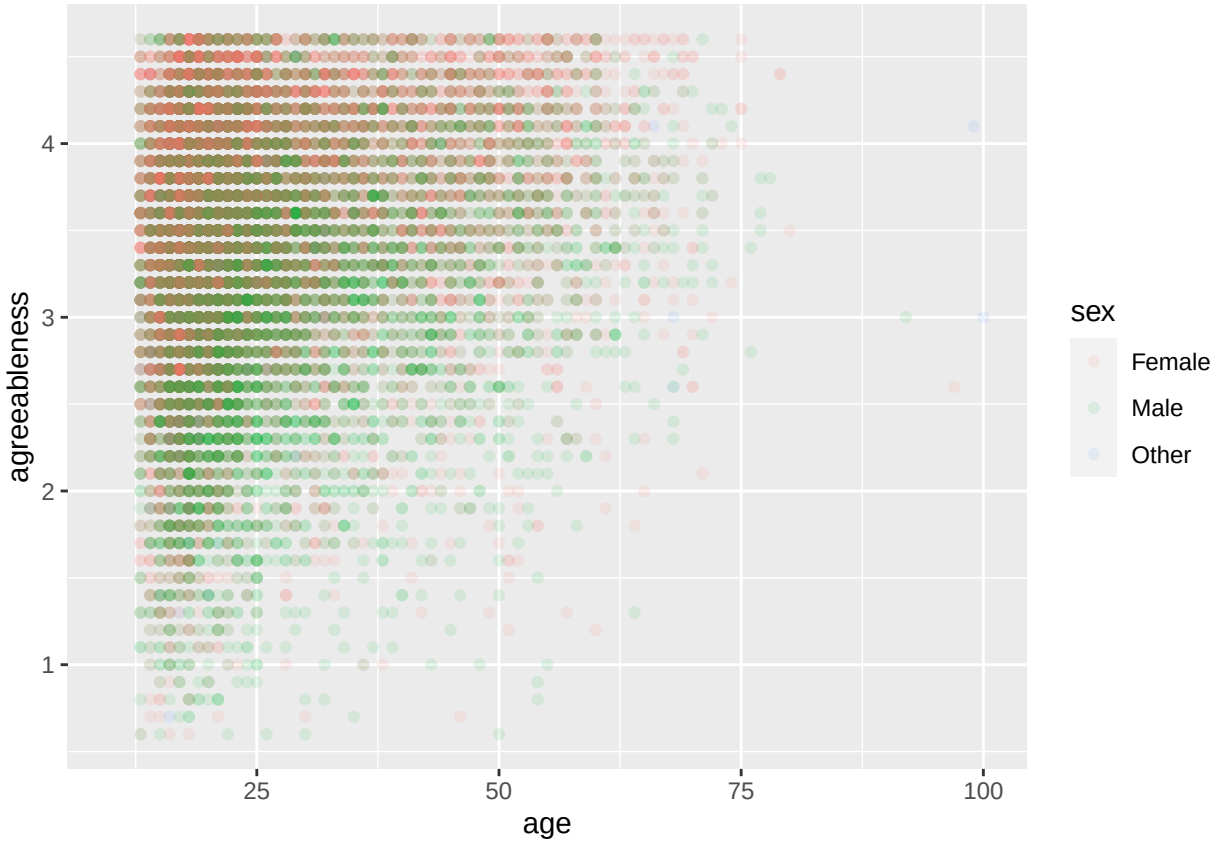

```
## Warning: Removed 83 rows containing missing values (geom_point).
```




```
## Warning: Removed 83 rows containing missing values (geom_point).
```




```
## Warning: Removed 83 rows containing missing values (geom_point).
```



Topics From Class

(a) Git
(b) RMarkdown
(c) Statistical concepts such as normal distributions, mean, standard deviations, percentiles, and areas under the curve
(d) Geometric distributions
(e) T tests to determine if two populations are significantly different.

Conclusion