Written Report on Factors that Help Diagnose Autism

The Repos - Jeffrey Bohrer, Alexandra Green, Anna Zhang, Kevin Lee

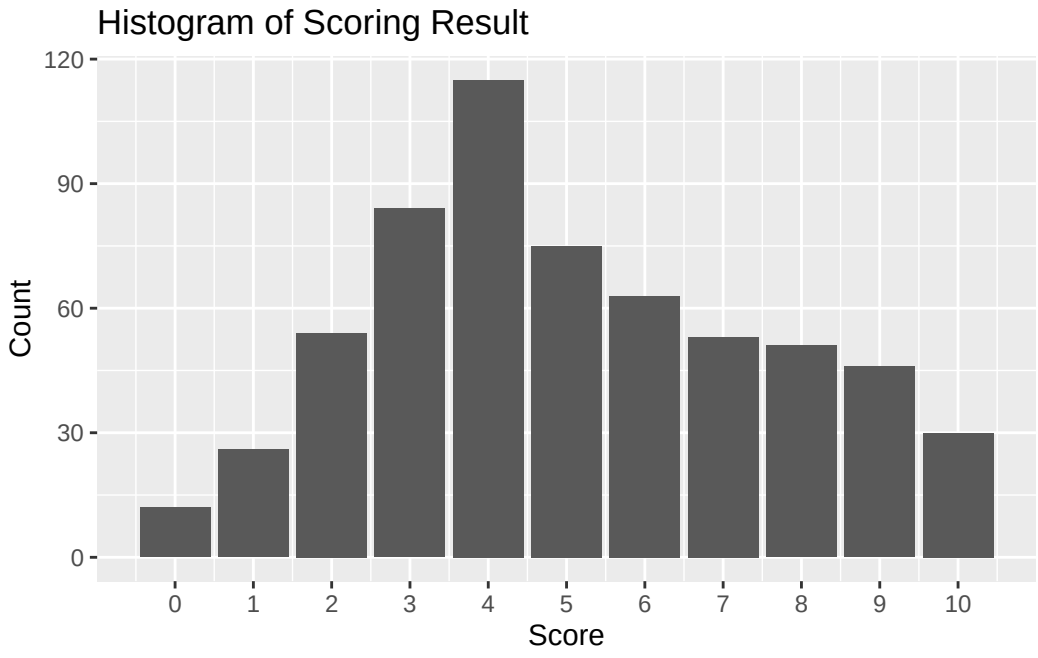
2025-03-20

Exploratory data analysis

Introduction

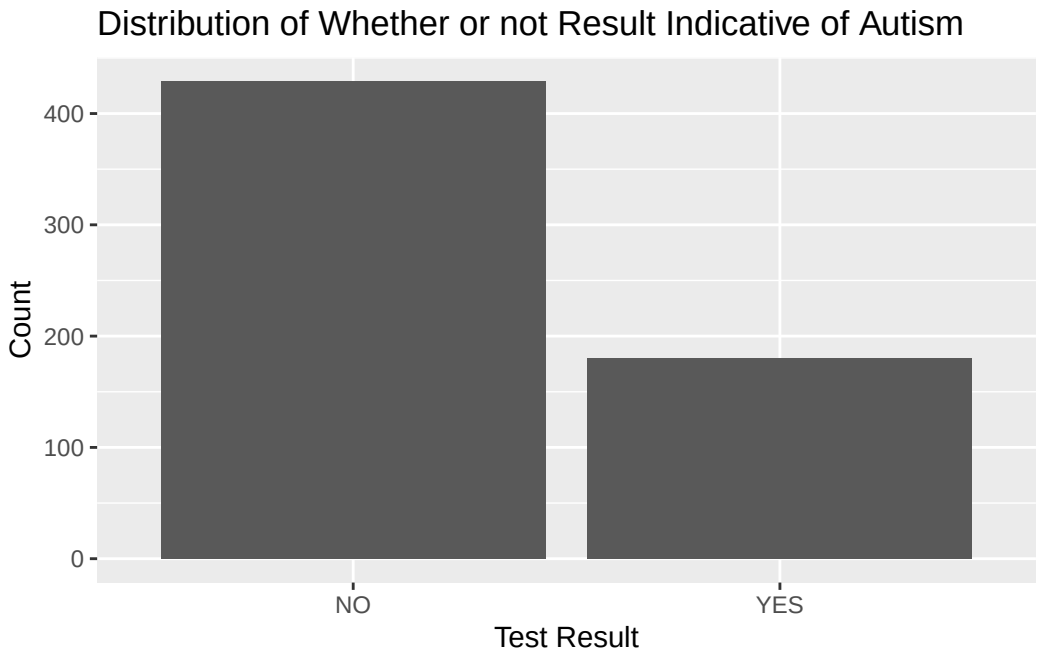
Univariate EDA

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   3.000   5.000   5.077   7.000  10.000 
```

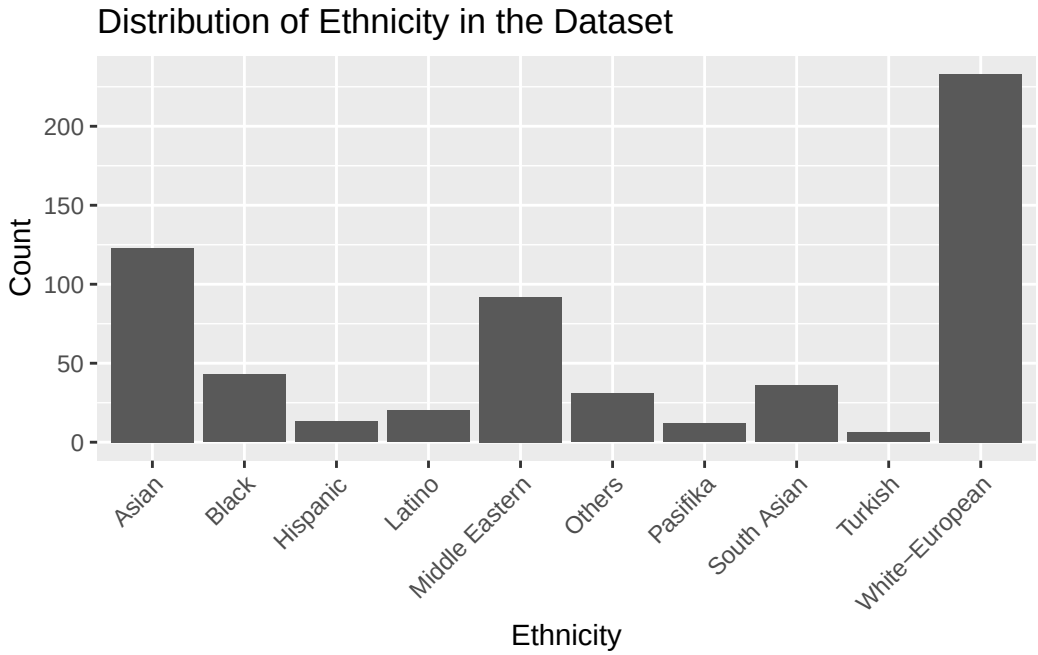


```
[1] 0.2955665
```

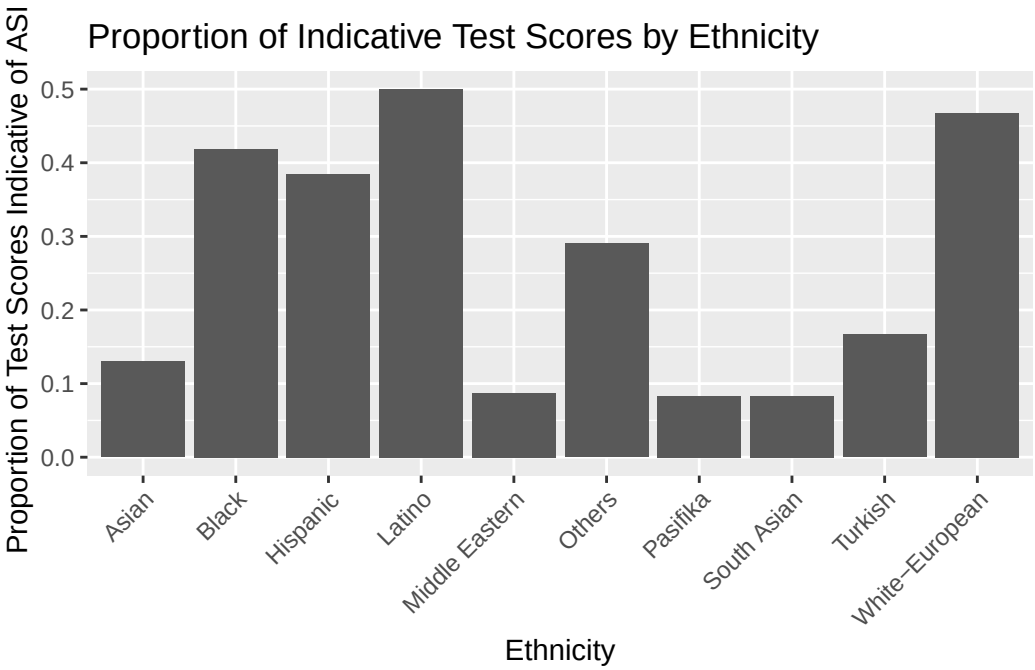
As we can see as a quick summary of our data, the variable we are most interested in – the final score of users – ranges from 0 to 10, which makes sense as there are 10 questions that can be answered either 'Yes' or 'No'. Generally, the scores are a bit right-skewed, with a single peak around 4 or 5 points. The mean is found to be 5.077 points, and the median is 5 points, which is relatively high given that a score of higher than 6 warrants further investigation into a diagnosis. However, as suspecting a diagnosis is reason to take a test in the first place, that is a reasonable reflection of the population taking this test. We then observe that roughly 30% of test takers are encouraged to seek a diagnosis, due to a score higher than 6. We also observe the IQR to be 4 points, as most test takers have a score between 3 and 7 points inclusive. This is a reasonable spread in the context of the potential values, and in this context, no outliers exist.



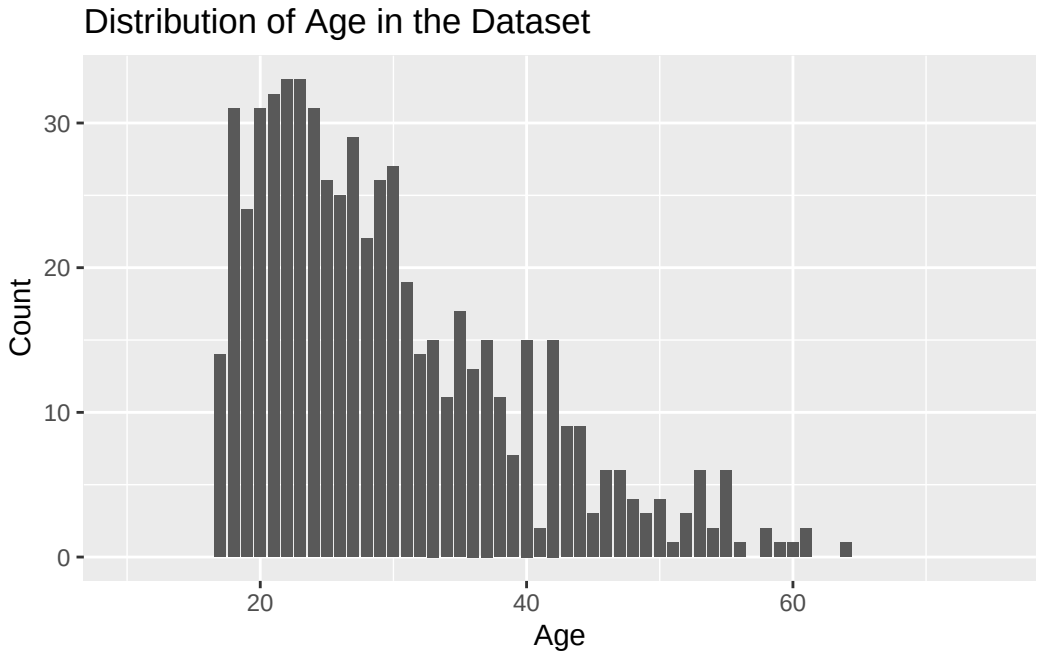
Additionally, here is a visual demonstration of the number of test takers receiving further encouragement to pursue a professional diagnosis of ASD. Though still significantly less than those not encouraged, a notable proportion of test takers were informed they had traits or behaviors indicative of autism.



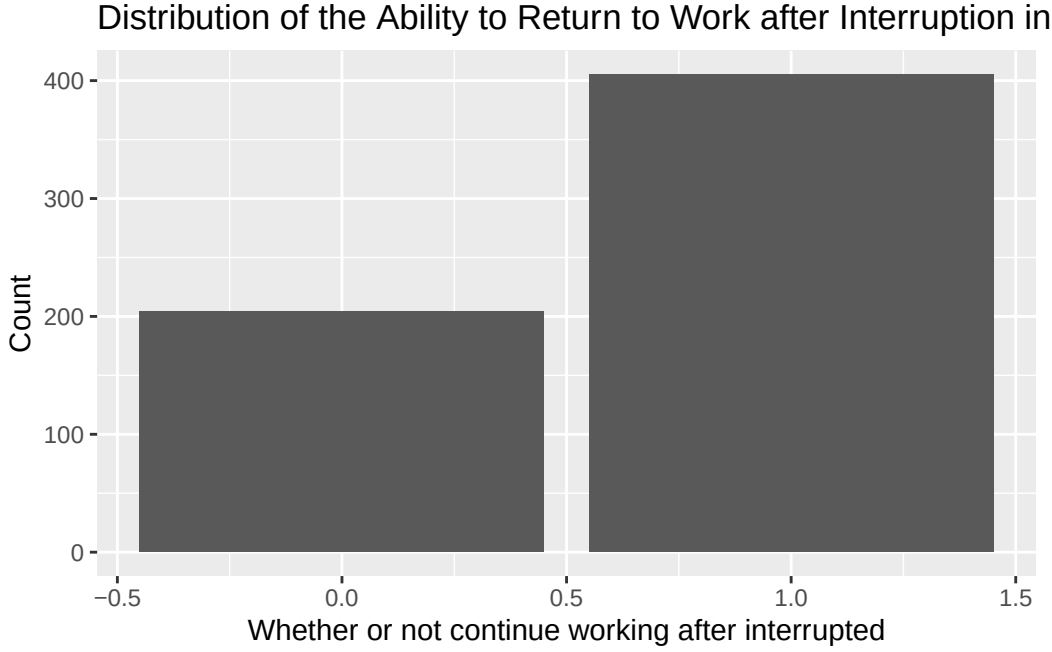
We observe here that the most common ethnicity was White-Europeans, with over 200 observations in our data set. Secondly and thirdly, we observe the Asian and Middle Eastern populations to be above 100 and slightly below 100 observations, respectively. All other ethnicities have less than 50 observations in our data set, meaning it may be more difficult to draw conclusions about these populations. We can now check and see how the percentages of being encouraged to seek a diagnosis differ by ethnicity:



As we can see here, the population with the highest proportion of test takers receiving a high test score was the Latino population, with close to half achieving higher than a 6. Since this is a comparably uncommon population in our data set, it does raise questions of whether it is indicative of the greater Latino population or merely due to the smaller sample size in our data set. Additionally, the White-European, Black, and Hispanic populations also demonstrate a greater proportion of high test scores. The variation in these proportions according to ethnicity warrants further exploration for how the distribution of test scores differs by ethnicity as well.



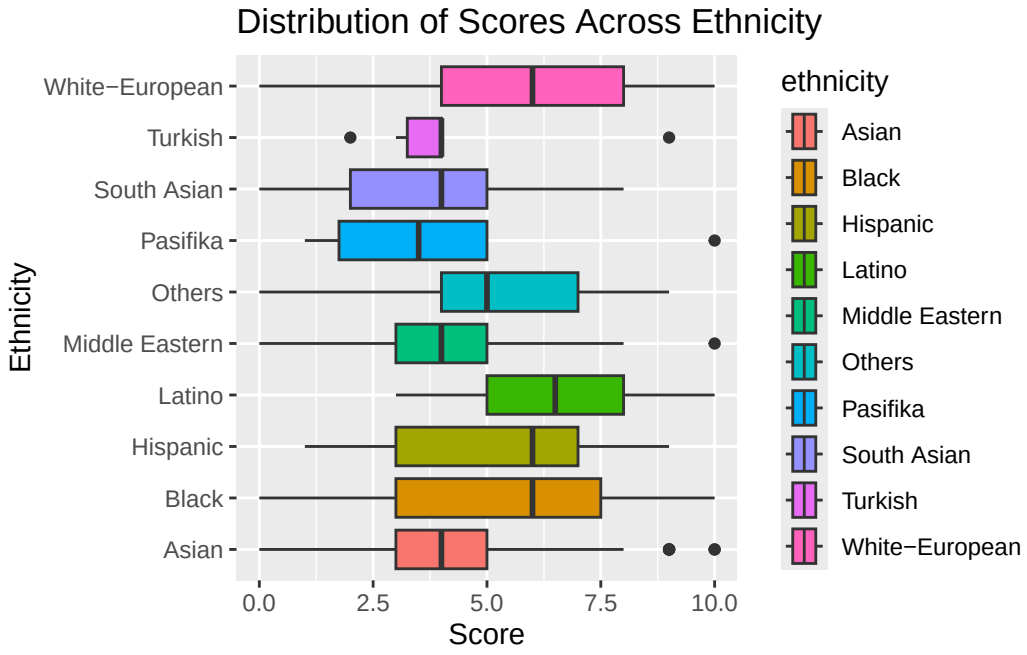

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 17.00   22.00   27.00   30.22   35.00  383.00
```

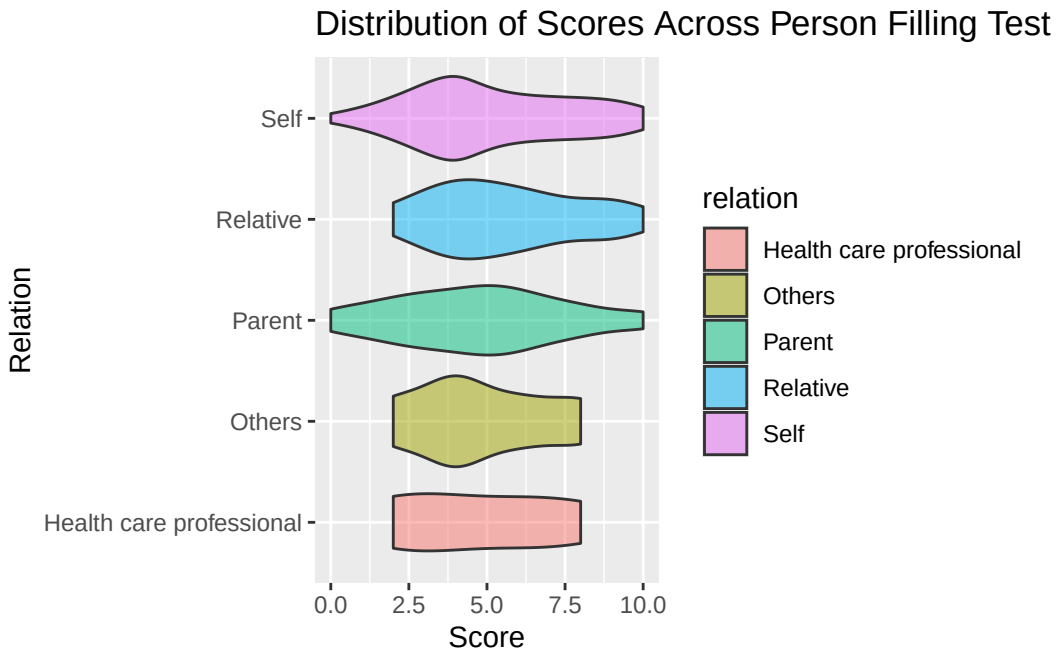


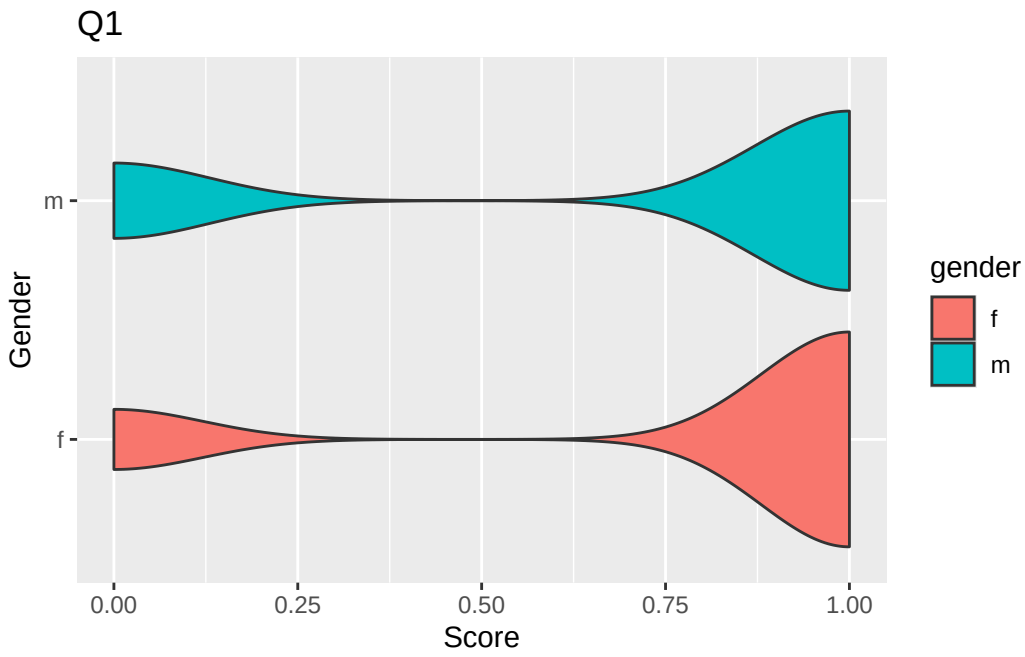
Generally, we observe the distribution of ages to possess a strong right-skewness, with a mean age of 30.22 and a median age of 27. There is a clear peak in the distribution roughly around early to mid 20's years of age. The ages possess a minimum of 17 up to a maximum age of 383, which is clearly a false observation that needs to be filtered from the data. The IQR is 13 years, which is a fairly small spread given the range of ages, as we observe that it seems the majority of test-takers are under 30 years old.

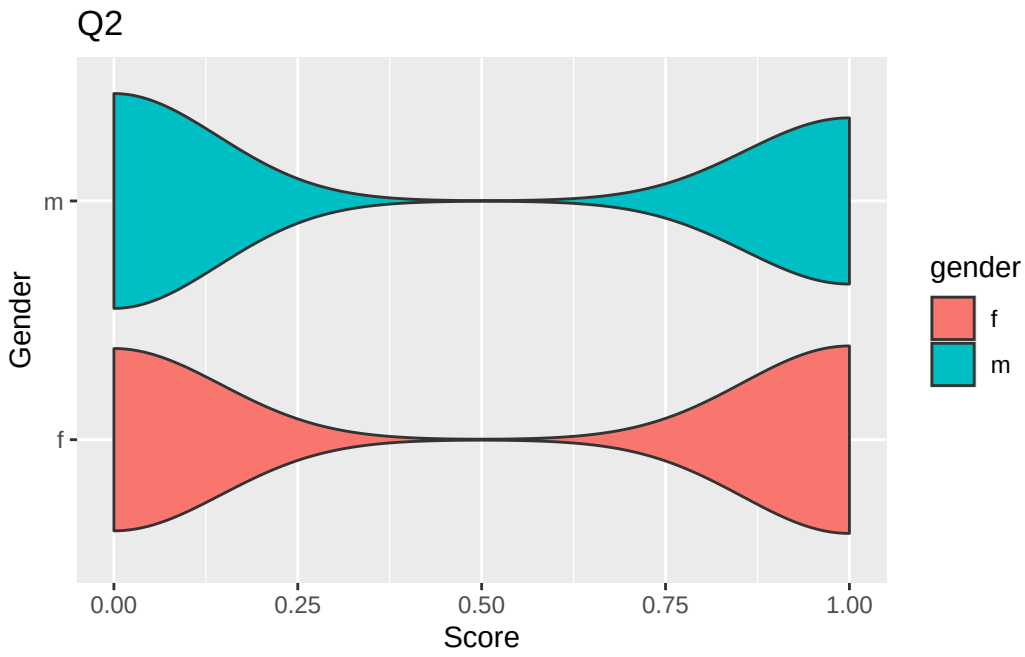
Additionally, as a brief glance at one of the ten questions on the test, we see that "the ability to continue working after an interruption" question had majority of test takers indicate that this was an area of difficulty, which acts as a potential indicator for ASD.

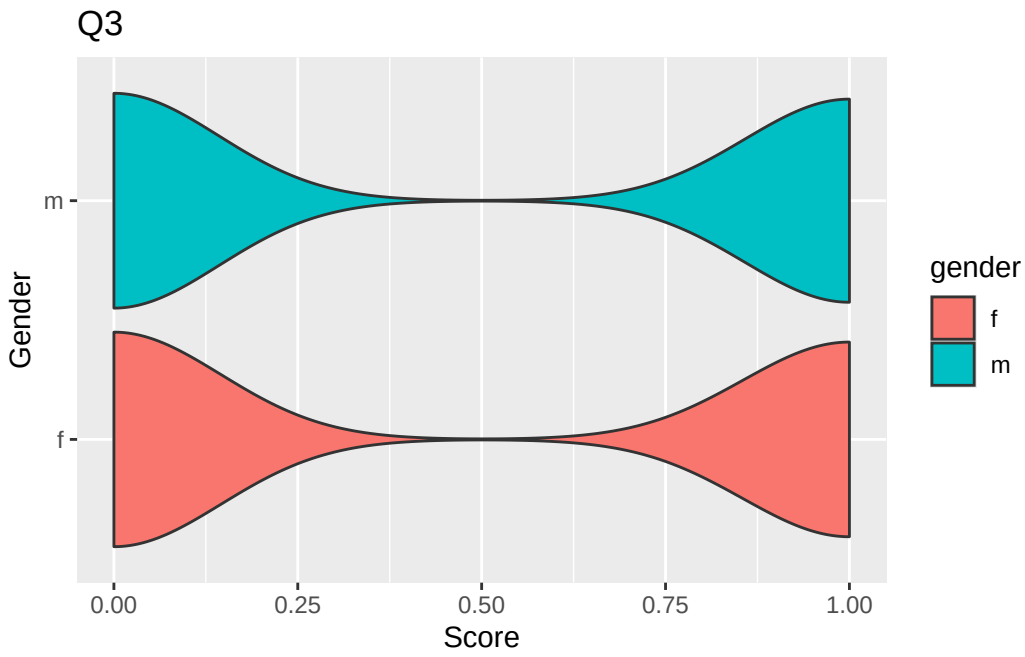
Bivariate EDA

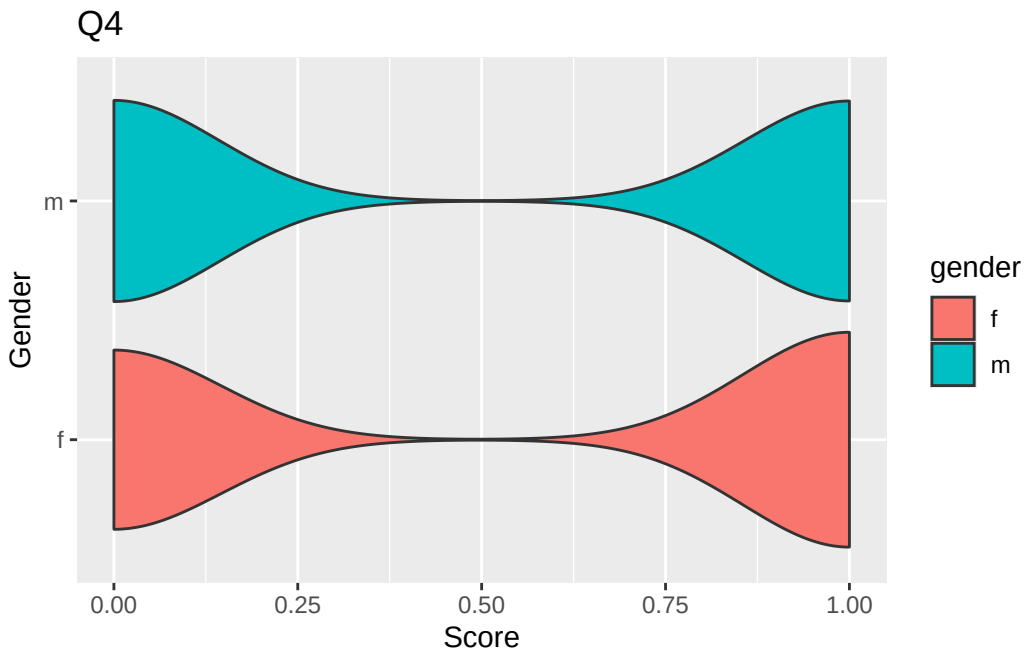


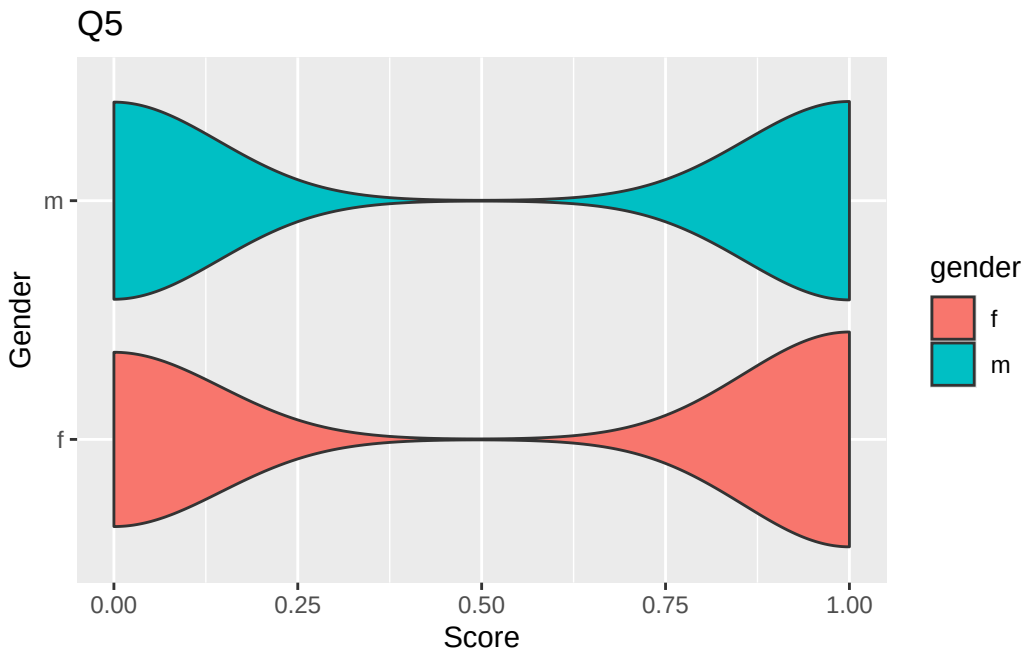

Distribution of Scores Across Person Filling Test
Relation
Self
Relative
Parent
Others
Health care professional
0.0
2.5
5.0
7.5
10.0
Score
relation
Health care professional
Others
Parent
Relative
Self

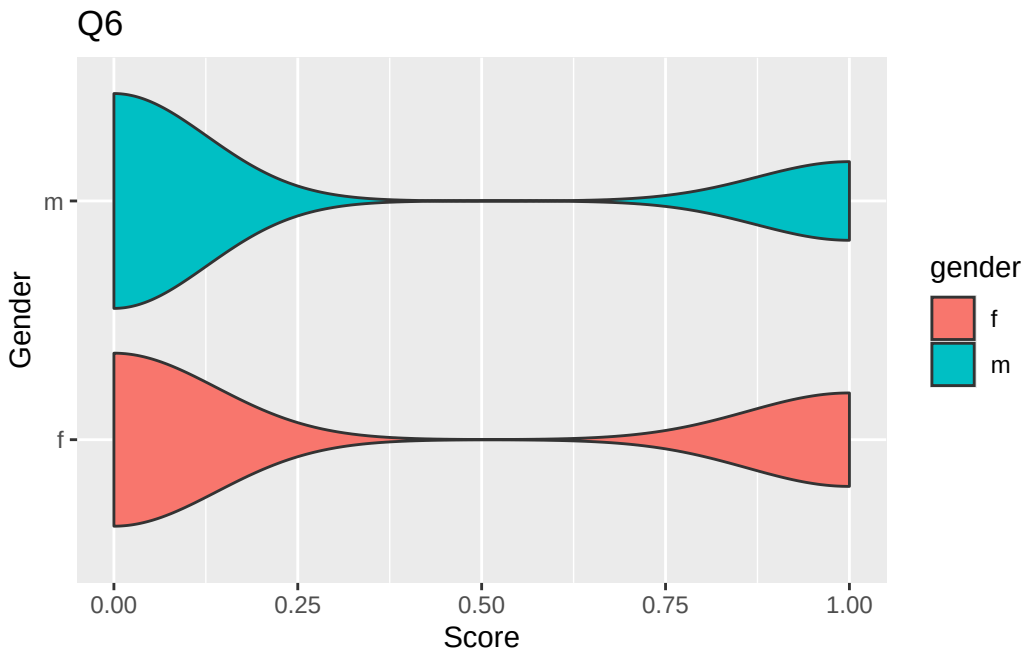

Q1
Gender
m
f
0.00
0.25
0.50
0.75
1.00
Score
gender
f
m

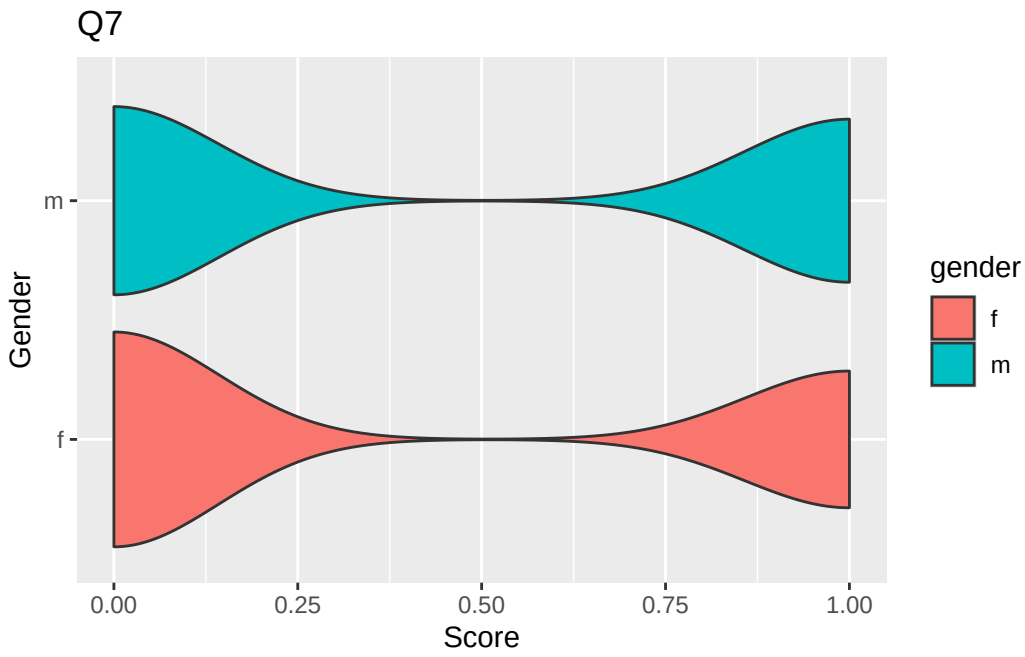

Q2
Gender
m
f
0.00
0.25
0.50
0.75
1.00
Score
gender
f
m

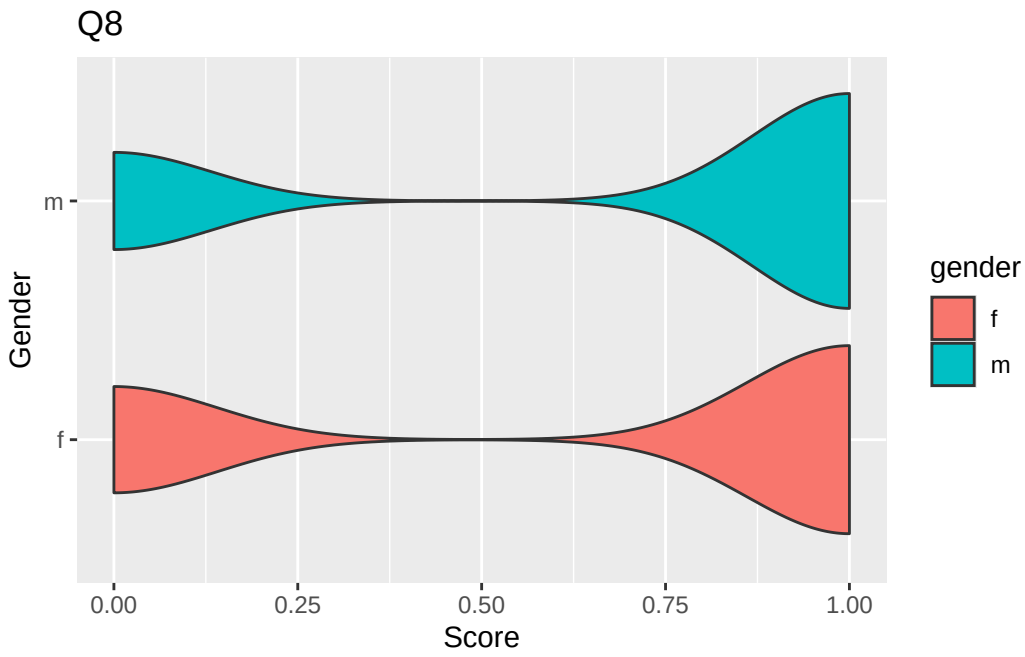

Q3
Gender
m
f
0.00
0.25
0.50
0.75
1.00
Score
gender
f
m

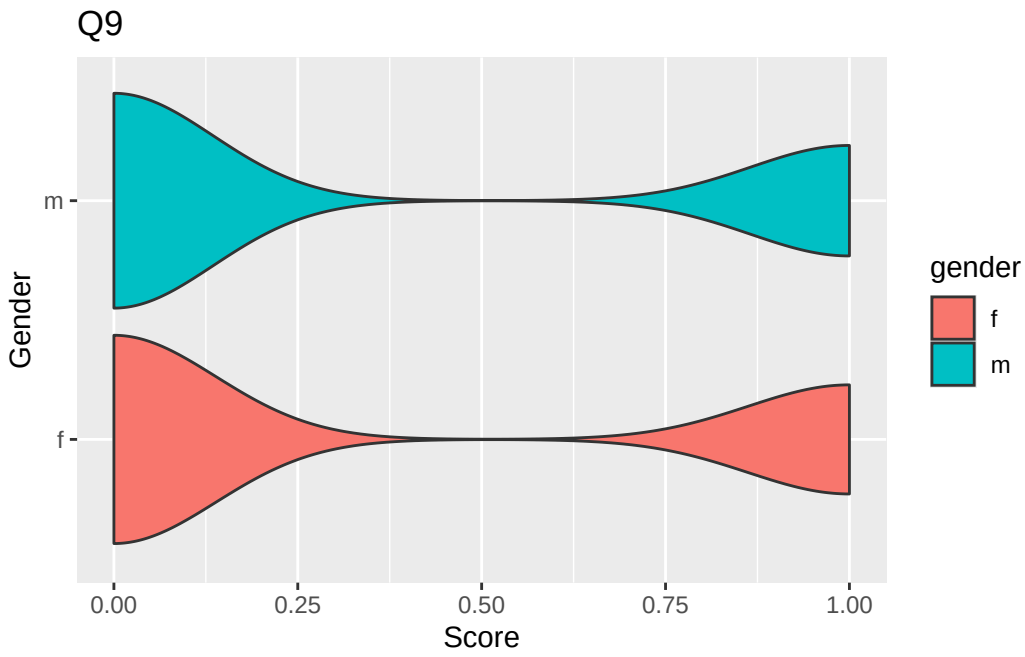

Q4
Gender
m
f
0.00
0.25
0.50
0.75
1.00
Score
gender
f
m

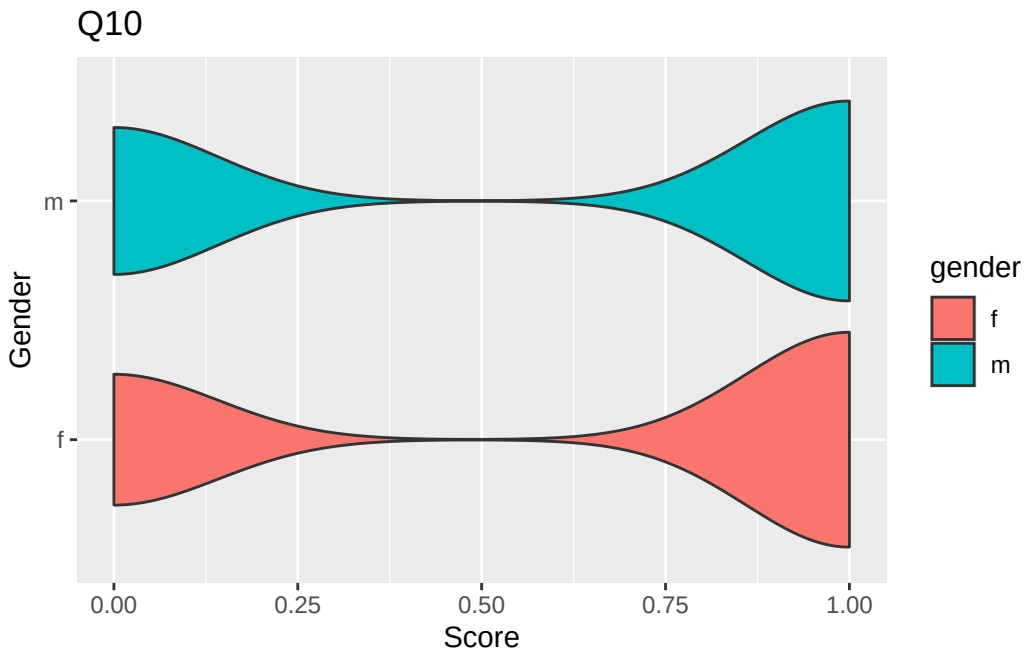

Q5
Gender
m
f
0.00
0.25
0.50
0.75
1.00
Score
gender
f
m

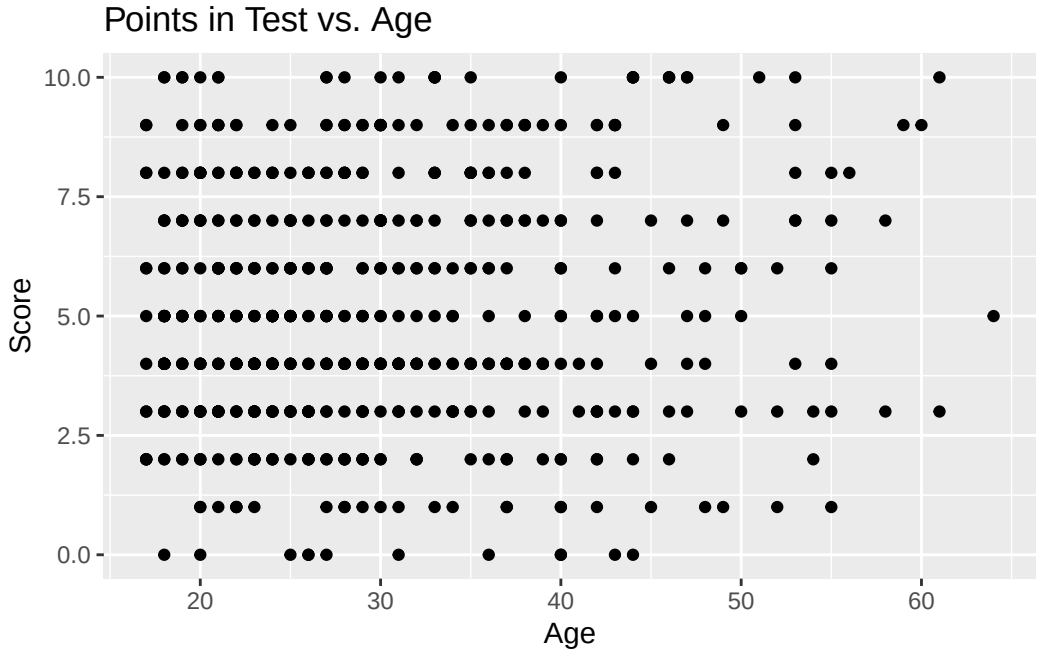

Q6
Gender
m
f
0.00
0.25
0.50
0.75
1.00
Score
gender
f
m


Q7
Gender
m
f
0.00
0.25
0.50
0.75
1.00
Score
gender
f
m


Q8
Gender
m
f
0.00
0.25
0.50
0.75
1.00
Score
gender
f
m


Q9
Gender
m
f
0.00
0.25
0.50
0.75
1.00
Score
gender
f
m





Here we present several graphs displaying bivariate relationships. With our first set of box plots, we find further evidence for our previous suspicion that test score distributions differ by ethnicity. Although most ethnicities have a median within the range of roughly 3 to 6, some ethnicities, like the White-European, Black, and Hispanic populations, demonstrate a

greater spread through their IQRs, while the Turkish, Middle Eastern, and Asian populations are much more concentrated around their medians. However, most ethnicities appear to have values almost entirely across the range of 0 to 10 in their test scores.

Interestingly, we can also observe how the relationship between the test taker and the subject of the questions may lead to different distributions of test scores. In the case when it is filled out by a relative or health care professional for example, the observed test score is 2 or greater, while in the case of the test taker being the subject or the parent of the subject, some still received a score of 0. This may reflect how personal biases or context affect truthfulness during the test.

Surprisingly, we can also observe minimal differences in how each gender answered the ten questions, with generally comparable distributions of 'Yes' and 'No'. This may be a point of further investigation as typically women tend to be underdiagnosed relative to men.

> **!** Important
>
> Before you submit, make sure your code chunks are turned off with `echo: false` and there are no warnings or messages with `warning: false` and `message: false` in the YAML.