EDA.Rmd

EDA Autism spectrum discorder quiz

Wouter Zeevat

Contents

Journal Thema 9 Wouter Zeevat

We will start off by looking at the data and codebook. The data contains 20 variables which will be loaded in as the codebook.

Variable.name.short	Variable.name.human.readable	type	unit
a1_score	Anwser question 1 score	numeric 1 or 0	score
a2_score	Anwser question 2 score	numeric 1 or 0	score
a3_score	Anwser question 3 score	numeric 1 or 0	score
a4_score	Anwser question 4 score	numeric 1 or 0	score
a5_score	Anwser question 5 score	numeric 1 or 0	score
a6_score	Anwser question 6 score	numeric 1 or 0	score
a7_score	Anwser question 7 score	numeric 1 or 0	score
a8_score	Anwser question 8 score	numeric 1 or 0	score
a9_score	Anwser question 9 score	numeric 1 or 0	score
a10_score	Anwser question 10 score	numeric 1 or 0	score
age	age	numeric	years
gender	gender	nominal	male or female
ethnicity	ethnicity	nominal	type of ethnicity
jaundice	jaundice	boolean	yes or no
autism	autism	boolean	yes or no
country_of_r	country of residence	nominal	country name
used_app_before	used the app before	boolean	yes or no
end_score	final test score	numeric	0-10 score
age_desc	age descending	nominal factor	years
relation	relation user compared to person of interest	nominal	string of relationship
class_asd	family member has asd	boolean	yes or no

The data

This is the data that will be used in the following project. it contains various information about adults doing an autism test. The columns speak for themselves except for the first 10. These columns represent the anwsers of the following question list.

https://www.nice.org.uk/guidance/cg142/resources/autism-spectrum-quotient-aq10-test-pdf-186582493

This is a general question list and the questions do not really matter. Each question gives points for the selected anwser. The more points the people have, the more chance there is of them having ADS (Autism disorder spectrum).

Research question

How accurate can the AQ-10 test predict whether someone has the autism spectrum disorder? The goal of this research question is to find out if this autism spectrum disorder test actually works and predicts someone has it. This would involve machine learning by testing if the computer would find correlations and would be able to predict them actually having ASD

After knowing all this it's time to see if the data is right. The data is supposed to have 20 columns and 704 rows.

```
## [1] 21
```

```
## [1] 704
```

Checking the data

The data also needs to be checked of missing data (A row that’s missing certain values). The ones that are missing important data will be removed. This needs to be done in order to not mess everything up. For example if someone is missing an anwser of the quiz, their score will be messed up and invalid.

This code will check if there are invalid values in any column.

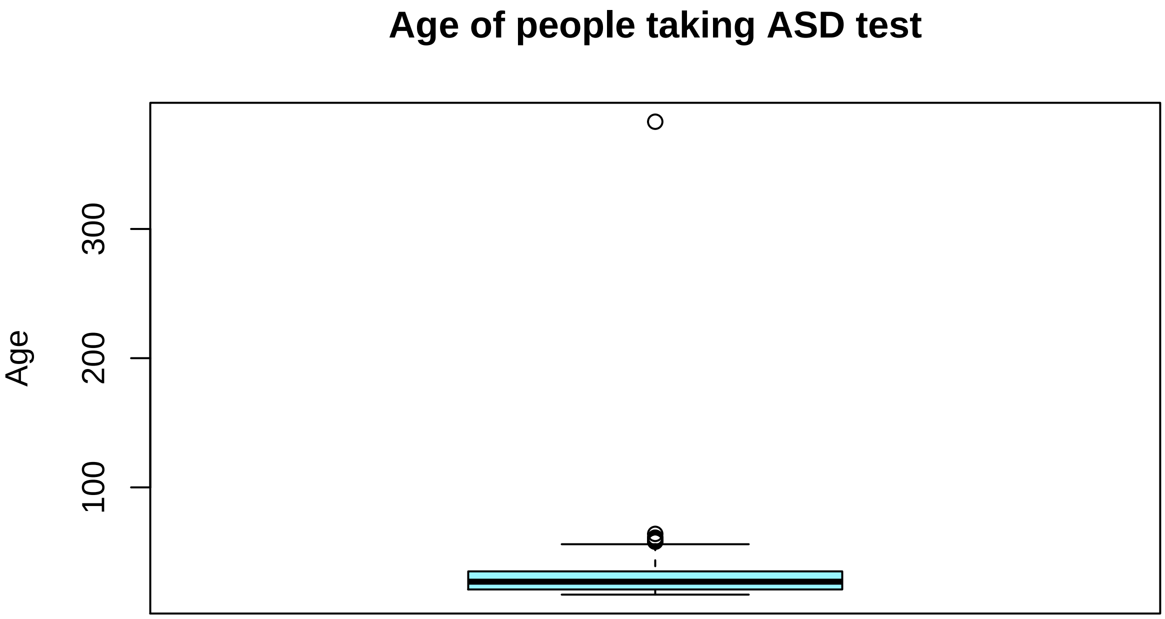
```
##    a1_score           a2_score           a3_score           a4_score
##  Length:704         Length:704         Length:704         Length:704
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    a5_score           a6_score           a7_score           a8_score
##  Length:704         Length:704         Length:704         Length:704
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    a9_score          a10_score              age           gender
##  Length:704         Length:704         Min.   : 17.0   Length:704
##  Class :character   Class :character   1st Qu.: 21.0   Class :character
##  Mode  :character   Mode  :character   Median : 27.0   Mode  :character
##                                        Mean   : 29.7
##                                        3rd Qu.: 35.0
##                                        Max.   :383.0
##                                        NA’s   :2
##   ethnicity           jaundice            autism          country_of_r
##  Length:704         Length:704         Length:704         Length:704
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  used_app_before      end_score       age_desc           relation
##  Length:704         Min.   : 0.000   Length:704         Length:704
##  Class :character   1st Qu.: 3.000   Class :character   Class :character
##  Mode  :character   Median : 4.000   Mode  :character   Mode  :character
##                     Mean   : 4.875
##                     3rd Qu.: 7.000
##                     Max.   :10.000
##
##   class_asd
##  Length:704
##  Class :character
##  Mode  :character
##
##
```

```
##
##
```

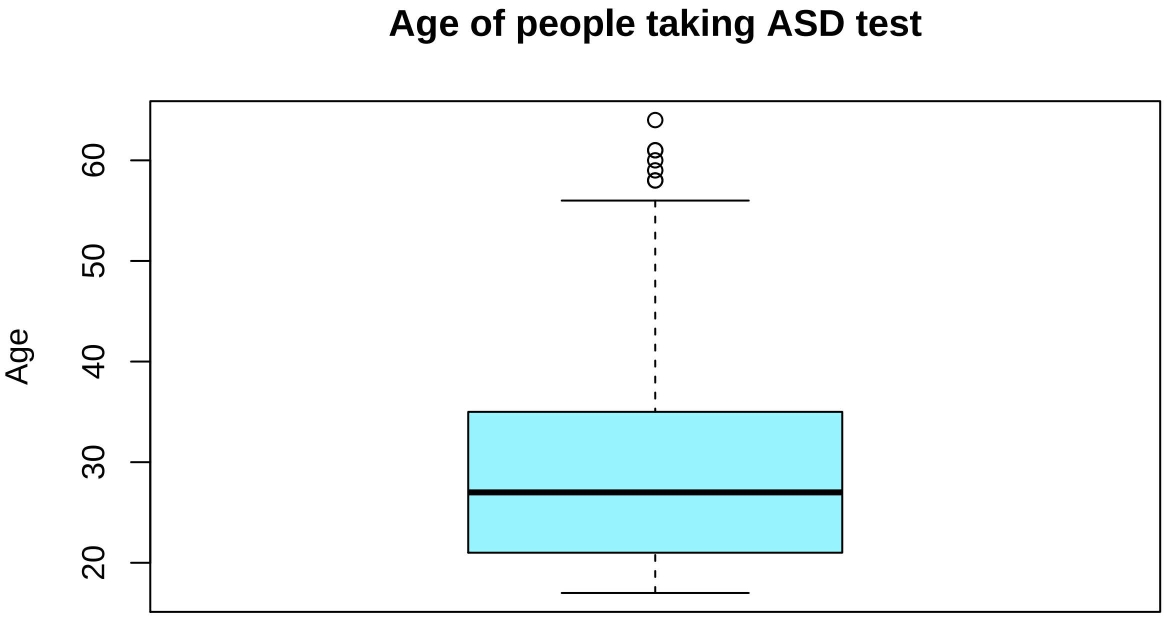
There are two NA's in the data. It is important to remove those in order to keep the data balanced. This will be done by removing their rows.

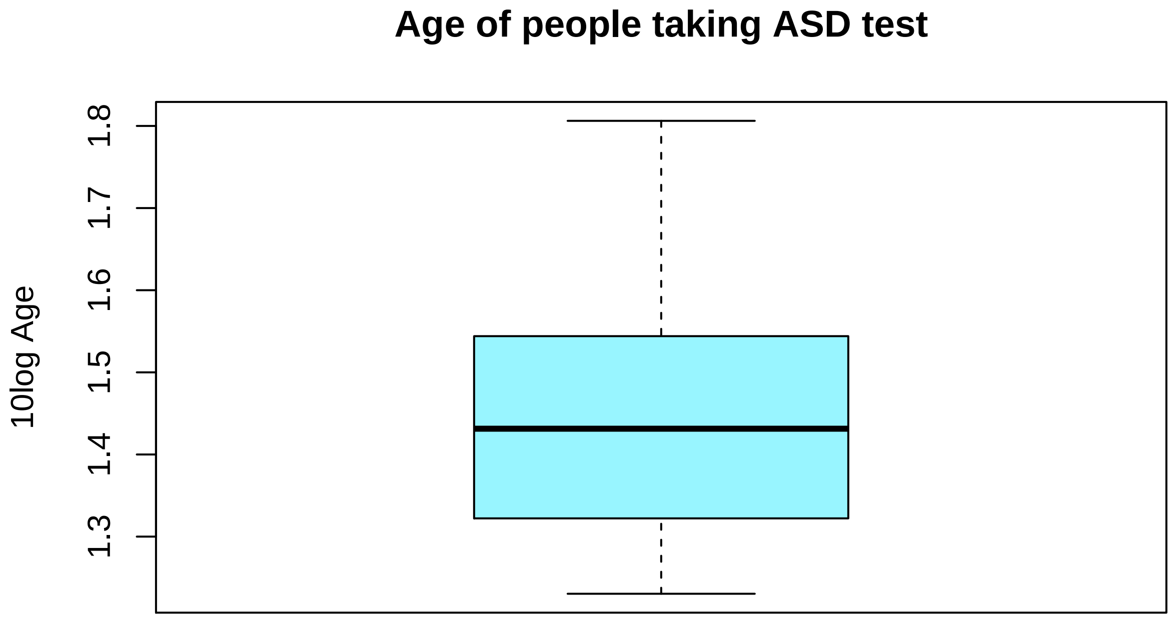
We will now take a look at the ages of the people taking the test are.

```
boxplot(data$age, main="Age of people taking ASD test", ylab="Age", col="cadetblue1")
```



As the boxplot shows, there's one huge outlier. One person would be 383 years old which just isn't humanly possible. The solution to this is taking out the whole row.

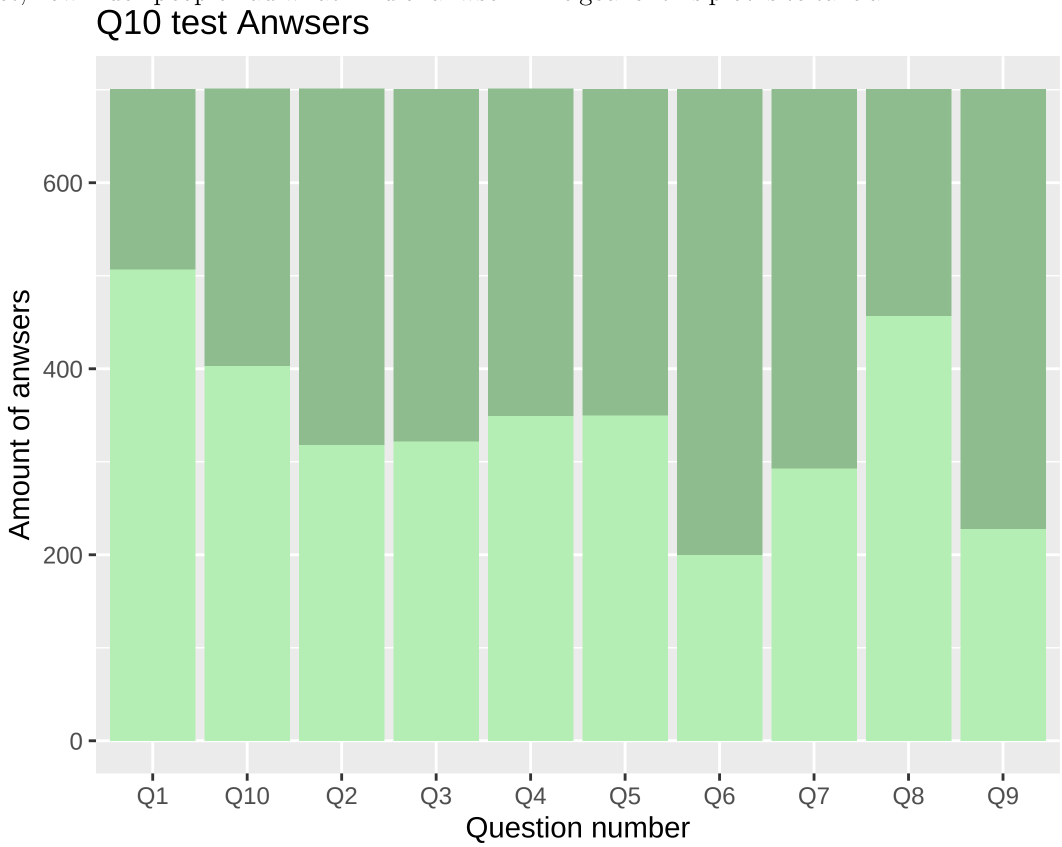




Correlations

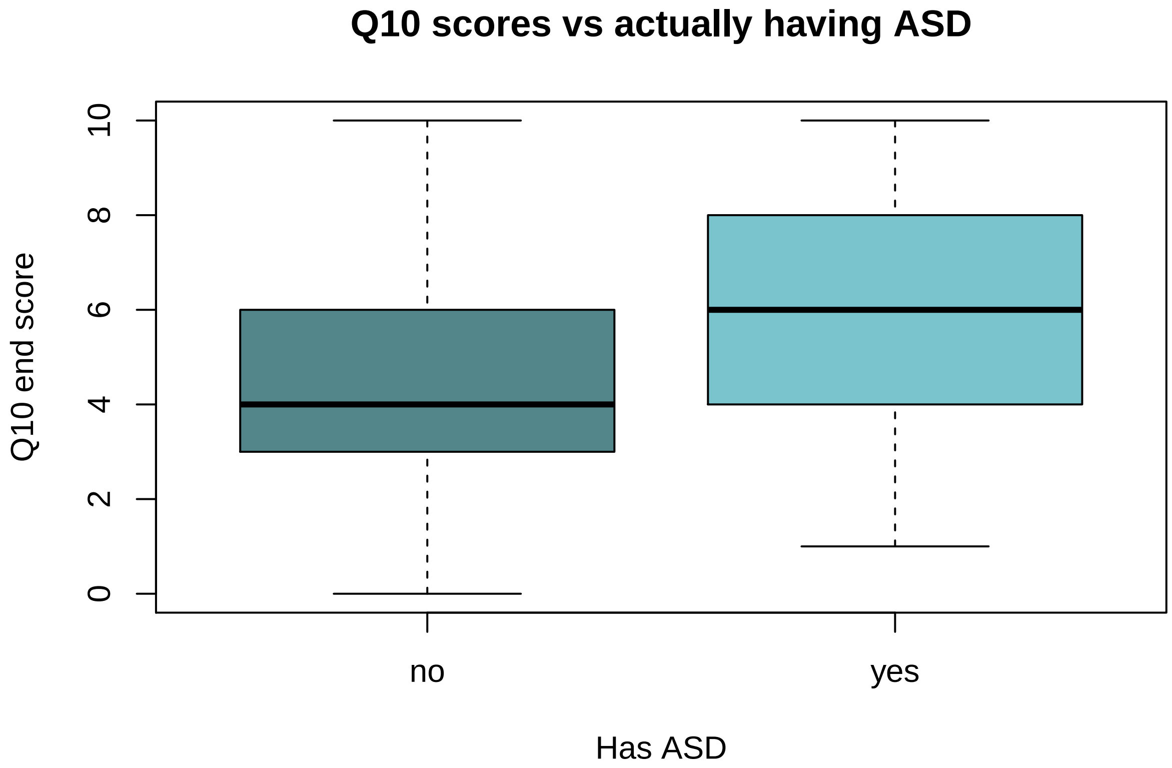
This boxplot shows that there's not much old people (60+) doing the test. The people who take this test are usually mid aged.

Now we will take a look at the test, how much people had what kind of anwser. The goal of this plot is to take a look at what the people scored.



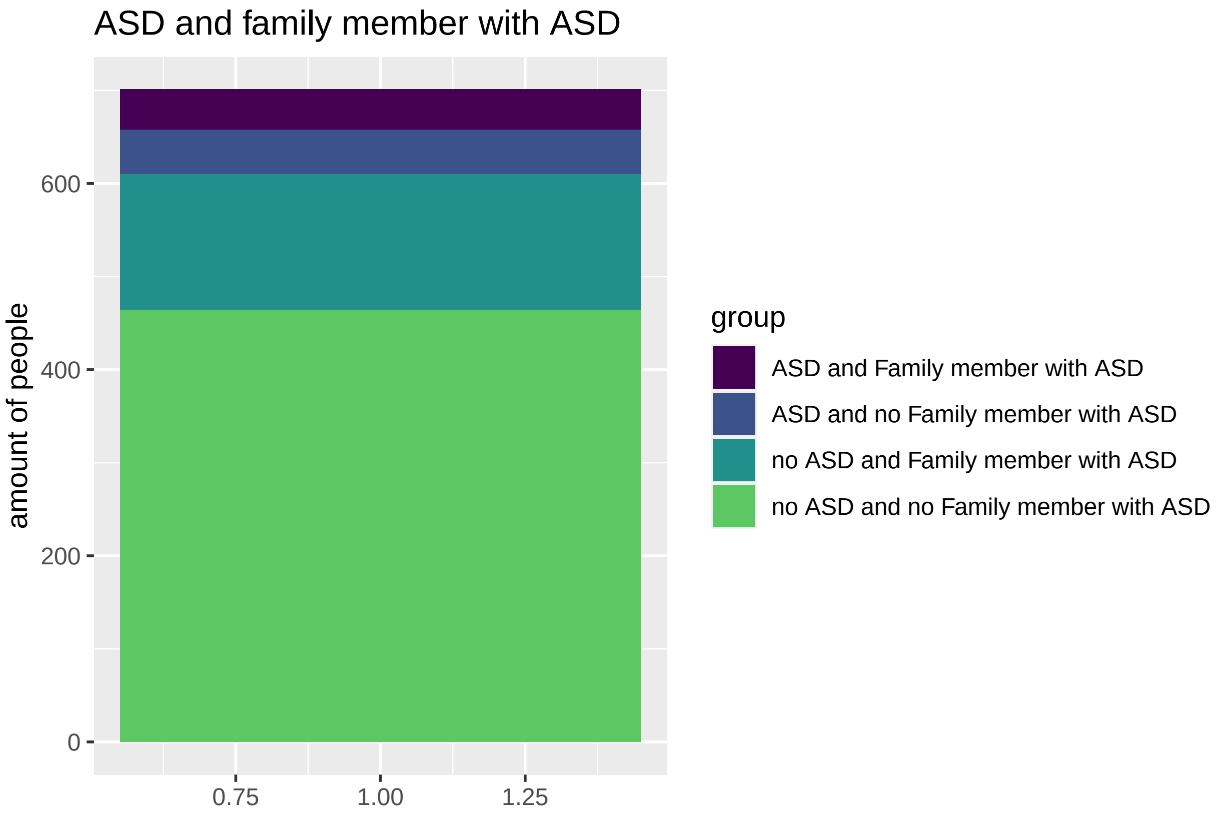
The conclusion of this plot is that most questions are anwsered positively (Without getting a point).

Let’s take a look at the correlations now. To start off, the end score will be measured against people actually having ASD. This will give a good view of the test because the test results will directly be compared to them having ASD.

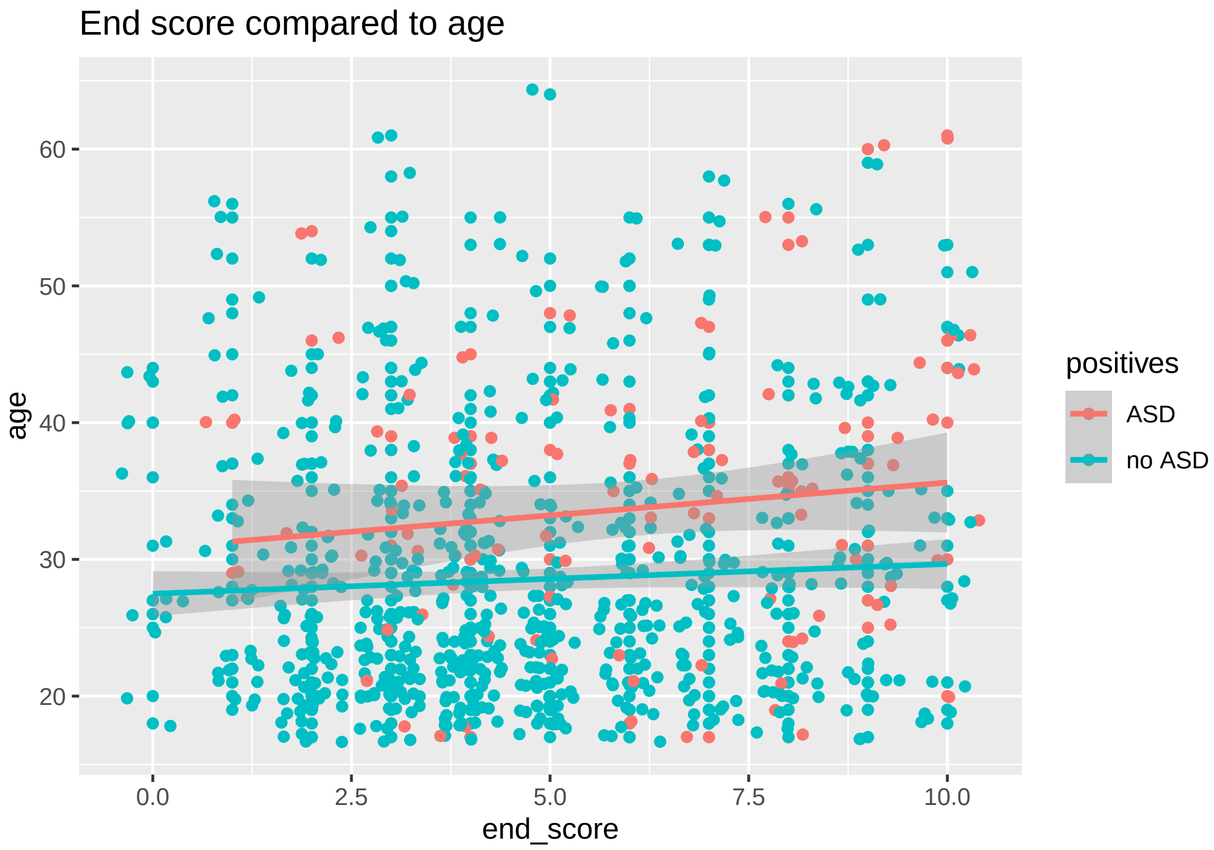


As seen in the plot, the scores do actually correlate with someone having ASD. This is true because the scores of the people having ASD are significantly higher than the other people. Let’s confirm this by doing a t-test

```
##
##  One Sample t-test
##
## data:  data$end_score
## t = 51.867, df = 700, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.703675 5.073785
## sample estimates:
## mean of x
##   4.88873
```



```
## ‘geom_smooth()‘ using formula ’y ~ x’
```


End score compared to age
age
60
50
40
30
20
0.0
2.5
5.0
7.5
10.0
end_score
positives
ASD
no ASD