Influence of Gender on Autism Spectrum Disorder (ASD)

Nina Kumagai and Yanan Cheah

1. Introduction: Problem Statement and Background

Autistic Spectrum Disorder (ASD) is a mental illness that limits an individual's linguistic, cognitive and social skills (Johnson & Myers, 2007). Behavioral and neuroimaging studies has consistently shown that ASD usually manifests differently in females, because females may have better social abilities than typical boys with ASD (Lai et al., 2015; Mandy et al., 2012). Because DSM metrics have been based mostly from data derived from male studies, current diagnostic methods (DSM-5) may overlook females with ASD (Kopp & Gillberg, 2011). There may also be a higher likelihood that females are diagnosed at a greater age, as symptoms become more pronounced at later life stage or as females become more self-aware of their characteristic symptoms (Howlin & Asgharian, 1999). Previous studies have not conducted an epidemiological analysis on gender-related autism regarding the new DSM-5 criteria (Newschaffer et al., 2007; Worley & Matson, 2012). To understand how females, respond differently to males, an analysis of how both genders respond to the questions in the new DSM-5 will be conducted.

The dataset is taken from the UCI website for the Centre of Machine Learning and Intelligent Systems. Originally, it was used for two main papers (Thabtah, 2017; Thabtah, 2018). In both studies, the data was mainly used to conduct analyses on how fulfilling current diagnostic methods are at determining the presence of ASD, in relation to the DSM-5. Machine learning was at the core of their analysis and was used to improve, precision, timing and quality of the diagnosis procedure, as well as to ensure all criteria (e.g. place of birth, presence of Jaundice) were significant for diagnosis purposes. Overall, they found that some of the diagnostic methods used previously in relation to the DSM-4 were not as relevant to the DSM-5.

The current investigation will focus on implementing a complex exploratory and predictive model to identify the influence gender may play in ASD diagnosis

The aims are to:

(1) Determine whether fewer females than males have been diagnosed with ASD.

(2) To understand whether females score differently on the DSM-5 diagnostic criteria.

The hypotheses are:

(1) Females are more likely to be diagnosed as having ASD at a later age compared to males.

(2) Questions related to social deficit in the DSM-5 diagnostic criteria (namely Q1, Q2, Q3, Q8, Q9, as shown at the end of the document), will be more relevant to males diagnosed with ASD compared to females.

Questions naming Q1 to Q10 comprises as follows:

- Q1, Q2, Q3, Q8 and Q9 are questions related to social deficit. All these questions addressed autistic impairments such as non-verbal communication skills, emotions reciprocity, social relationships and expression in communication.
- Q4, Q5, Q6, Q7 and Q10 are those that are not related to social deficit.

2. Methods

Data Structure:

Datasets were originally in .arff format. The file was saved in “csv”" format and each dataset (child, adolescent and adult) were transferred to R Studio. The “rbind” function was then used to combine the datasets together into one large dataset called “autism”. An additional column was added to the large dataset to signify which life stages (adult, adolescent, child) everyone corresponded to. This helped to ease the process of sub-setting individuals into various age groups for further analysis.

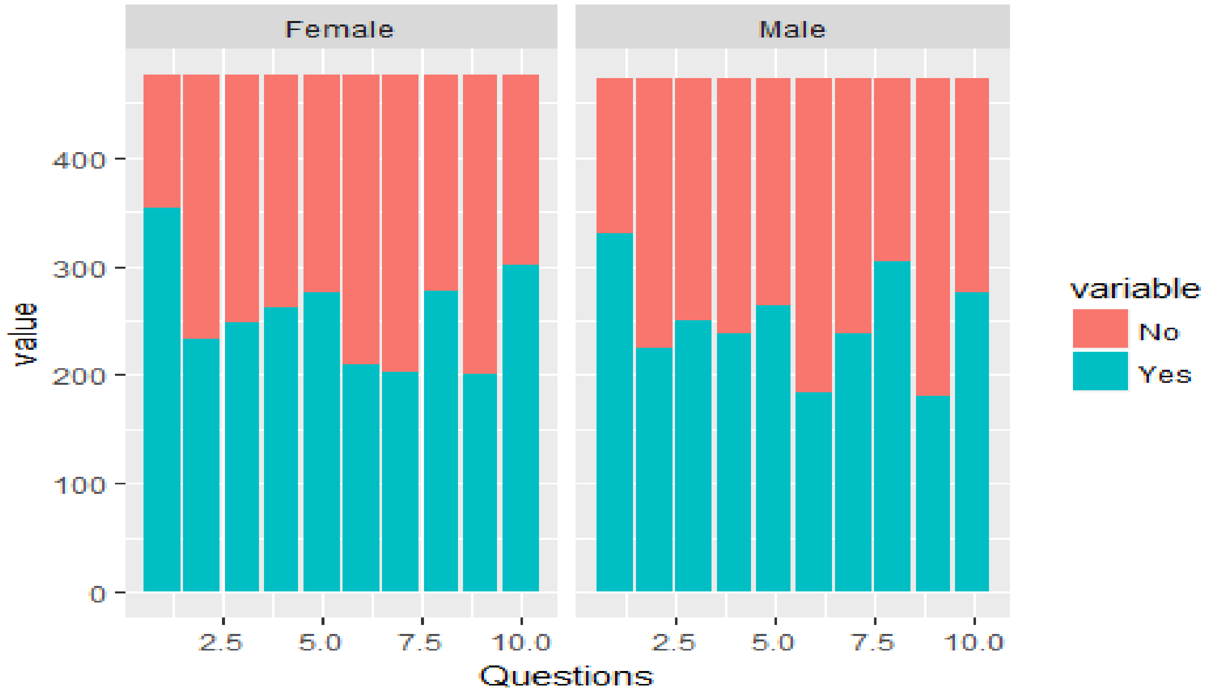
The data has a total of 948 people along with 22 variables. Data comprises of binary responses to 10 Questions, with 1 representing “yes”, and 0 representing “no”. Age, Gender, Race, Place of Residence, Age Range are also included, as character values. ASD column show categorical data on whether the individual had autism or not. Participants are only shown to have autism if they score seven or higher in Screening Score (they answered yes to at least 7 out of 10 questions within the DSM-5).

Data Cleaning and Wrangling:

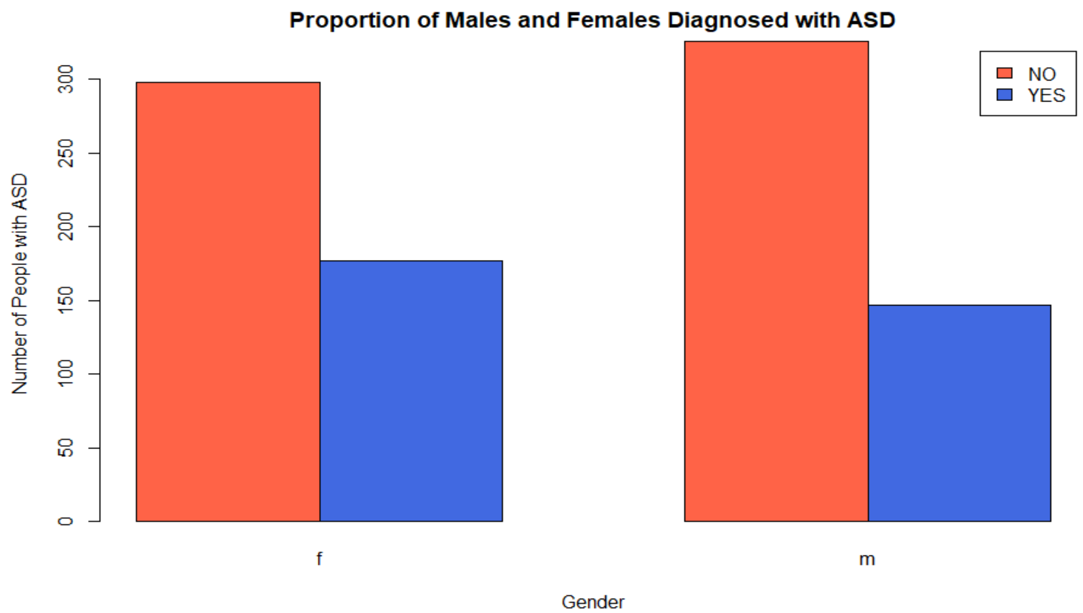
As the original dataset contained more males than females, approximately 150 males were randomly deleted from the dataset, making sure to maintain a similar proportional distribution of children, adolescents and adults to the female dataset. After that, the dataset was split based on gender, into two separate datasets.

The characteristics of this dataset are shown using the describe function. With the use of this function, missing values can be identified and deleted if necessary. Two outliers were also deleted from dataset. Additionally, the age variable that was originally a character value is subsequently set to an integer value. The binary response from those 10 Questions were changed into ‘yes’ and ‘no’ for ease of analysis and interpretation. They were also subsequently changed to a factor variable to enable random forest analysis.

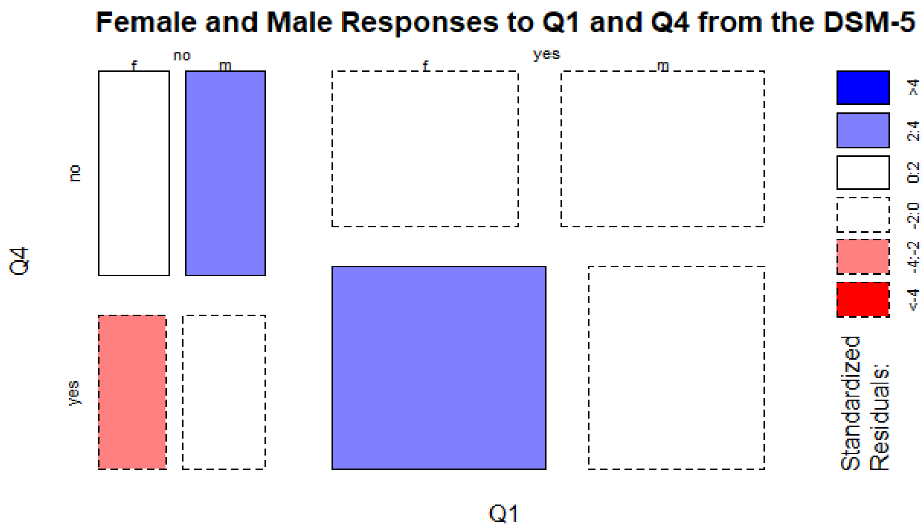
Exploratory Analysis:



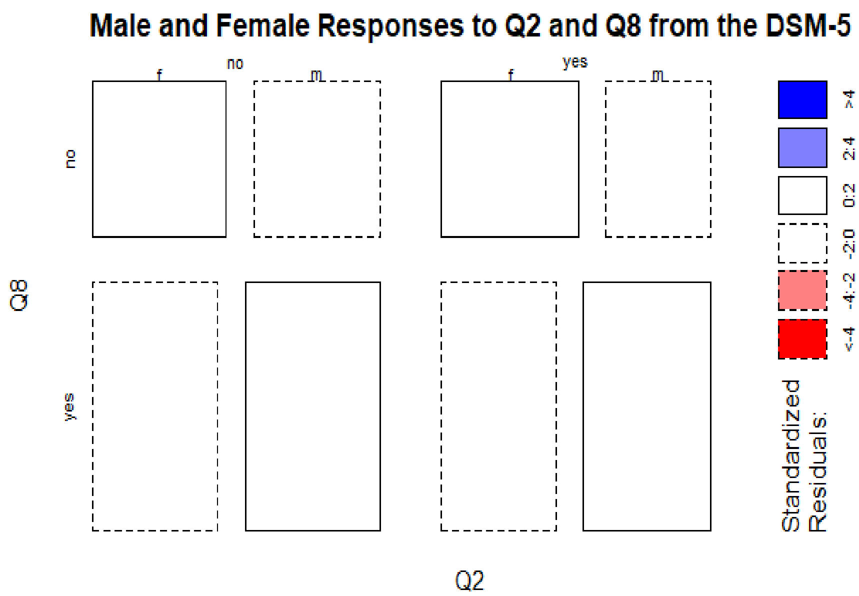
This plot uses the package ggplot2. The bar graph is plotted to address frequency of females and males responding yes to questions one to ten. The plot shows that there is a large variance in responses for each question and thus further analysis is desirable.



The plot shows a large variance in distribution of answers for each question between both genders. This bar plot indicates higher number of females that were diagnosed as having ASD than males. This indicates that our first hypothesis that less females will be diagnosed as having ASD compared to males may be incorrect. However, further analysis will be required to conclude this.



The plot shown is known as mosaic plot, which was plotted using vcd and vcdExtra package. This is done to compare 3 variables, which are gender, Q1 and Q2. The width of each box represents number of females or males answering yes or no to questions one and four. The plot indicates that slightly more females are likely to answer yes in Q1 and yes in Q4. As Q1 and Q4 conveys is in regards to social deficit and not social deficit respectively, there is some reason to suspect that the second hypothesis may not be valid.



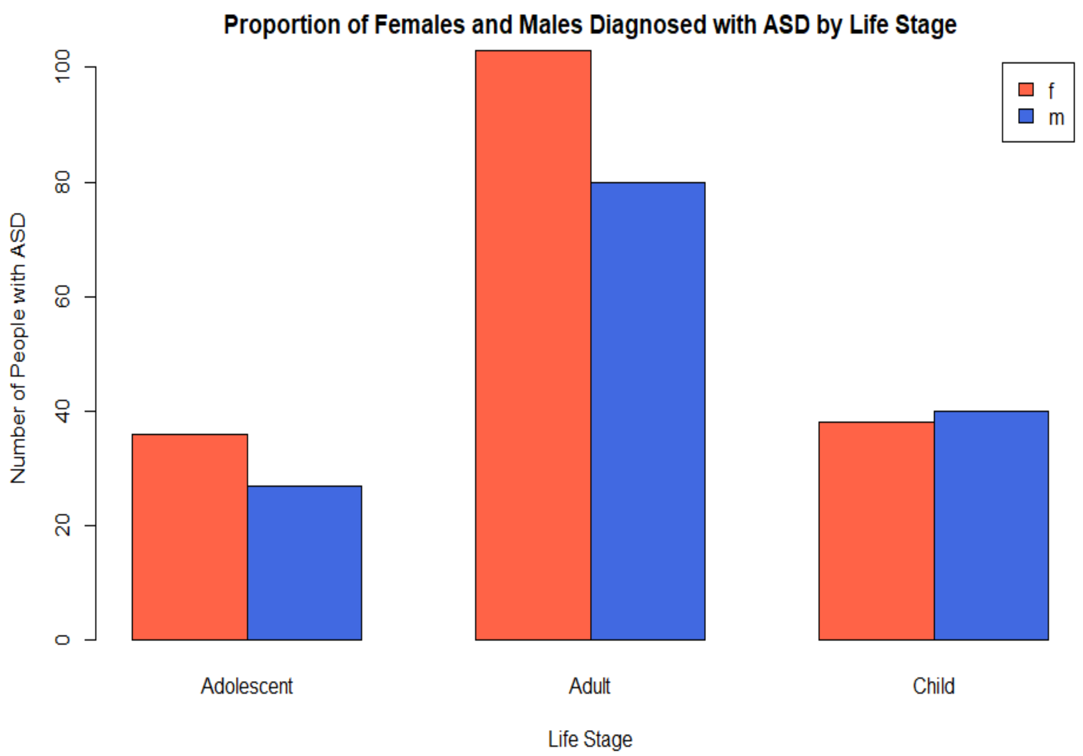
This mosaic plot does not show much variation in number of females and males when answering question two and question eight. This may show that questions related to social deficit may play an equally important role in both genders.

Predictive Analysis:

Decision trees and confusion matrix comprised the bulk of the analysis. This was also accompanied with a barplot for hypothesis 1 using the ggplot function. Other methods were applied (such as KNN and logistic regression) however both did not apply well to the binomial dataset that we had. In particular, KNN required the use of dummy variables and there was not sufficient time for such an analysis to be constructed.

3. Results

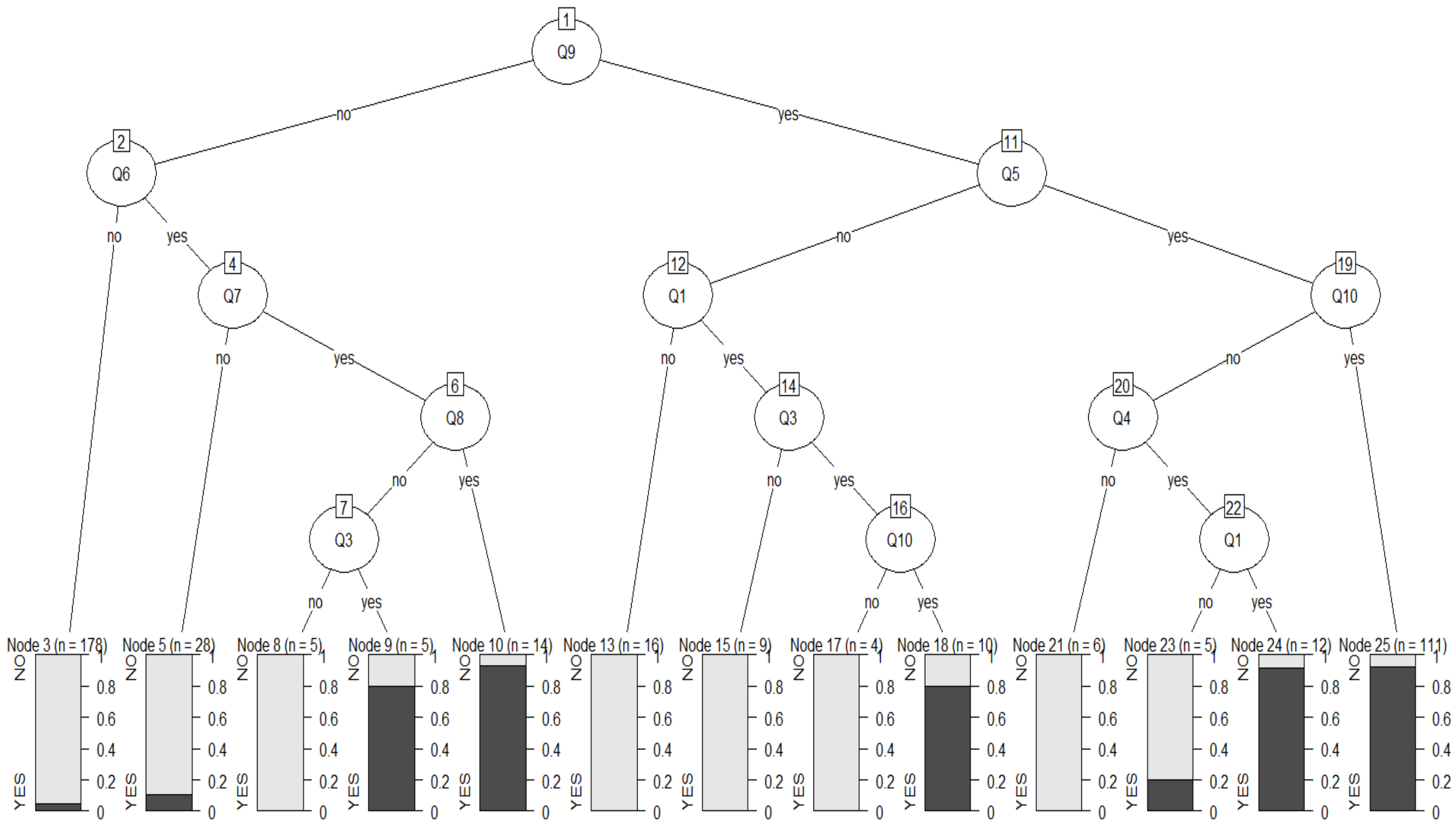
Hypothesis 1: Females are more likely to be diagnosed as having ASD at a later age *compared to males.*



The bar plot distribution shows larger proportion of females than males that are diagnosed with ASD within the adolescent and adult group. However, in the child group, there is a slightly larger number of males with ASD compared to females. Thus, it seems, females are more likely to be diagnosed as having ASD at a later age compared to males.

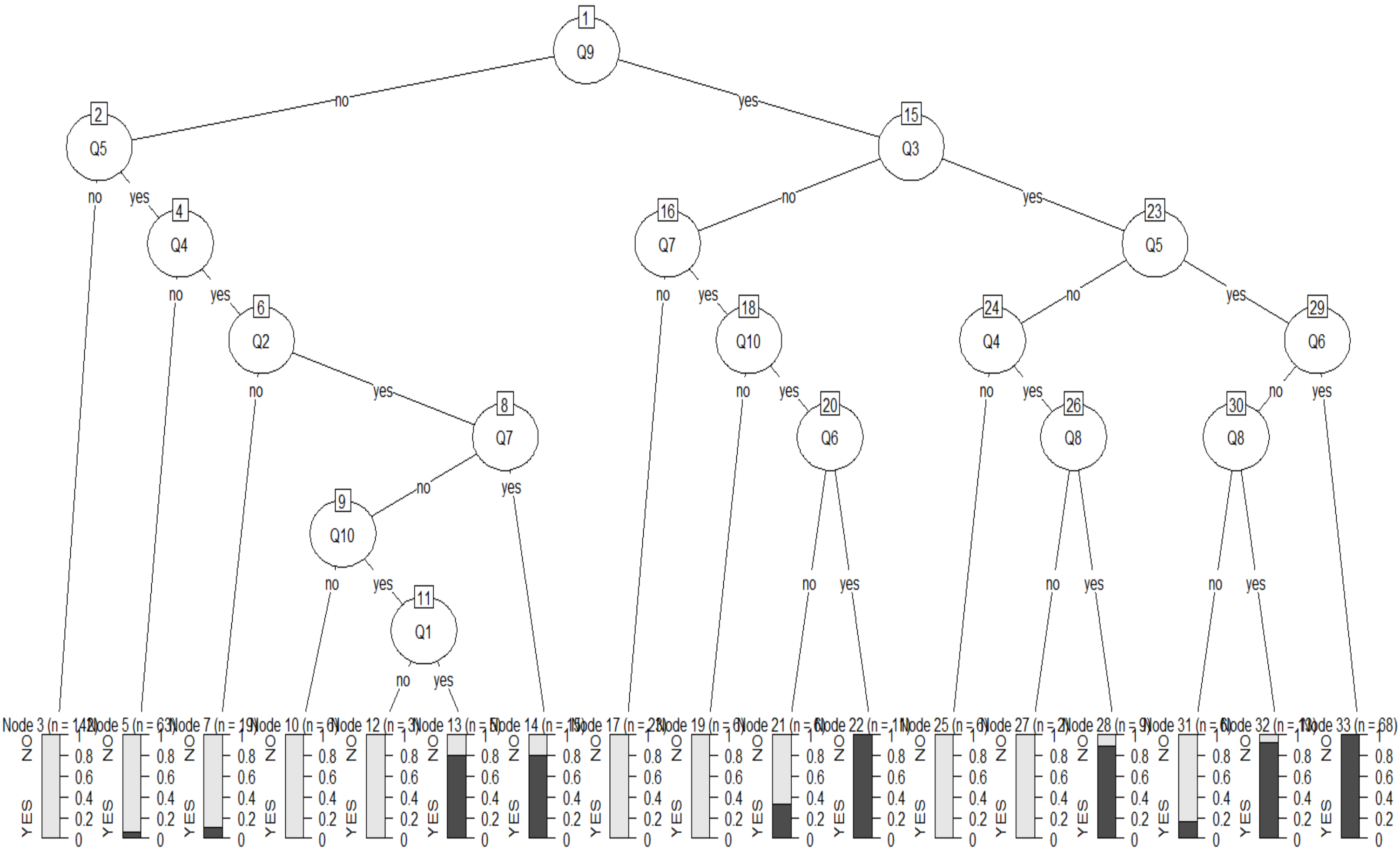
Hypothesis 2: Questions related to social deficit in the DSM-5 diagnostic criteria (namely Q1, Q2, Q3, Q8 and Q9) will be more relevant to males with ASD than females with the similar conditions.

Decision Tree for Females:



A decision tree shows relative importance of each question for each gender. The most important question for females is Q9, which is social deficit related. Followed by splitting via Q6 and Q5 which are both non-social deficit related. For example, if a female were to respond yes towards Q9, Q5 and Q10, this individual is more likely to be diagnosed as non-autistic.

<u>Decision Tree for Males:</u>



The decision tree for males involved more splitting of questions for diagnosis. This may indicate that more questions are significant towards diagnosing a male with ASD, compared to female. Nevertheless, Q9 is also the most important question for males, which is the same for females. Followed by splitting with Q5 and Q3. If an easy example is taken, a male need to answer yes to Q9, Q3, Q5 and Q6 to have more likelihood in being diagnosed without having ASD.

Confusion Matrix for Females:

```
           Actual
Predictions NO YES
        NO  40   4
        YES  1  26
```

Confusion Matrix for Males:

```
           Actual
Predictions NO YES
        NO  46   1
        YES  1  23
```

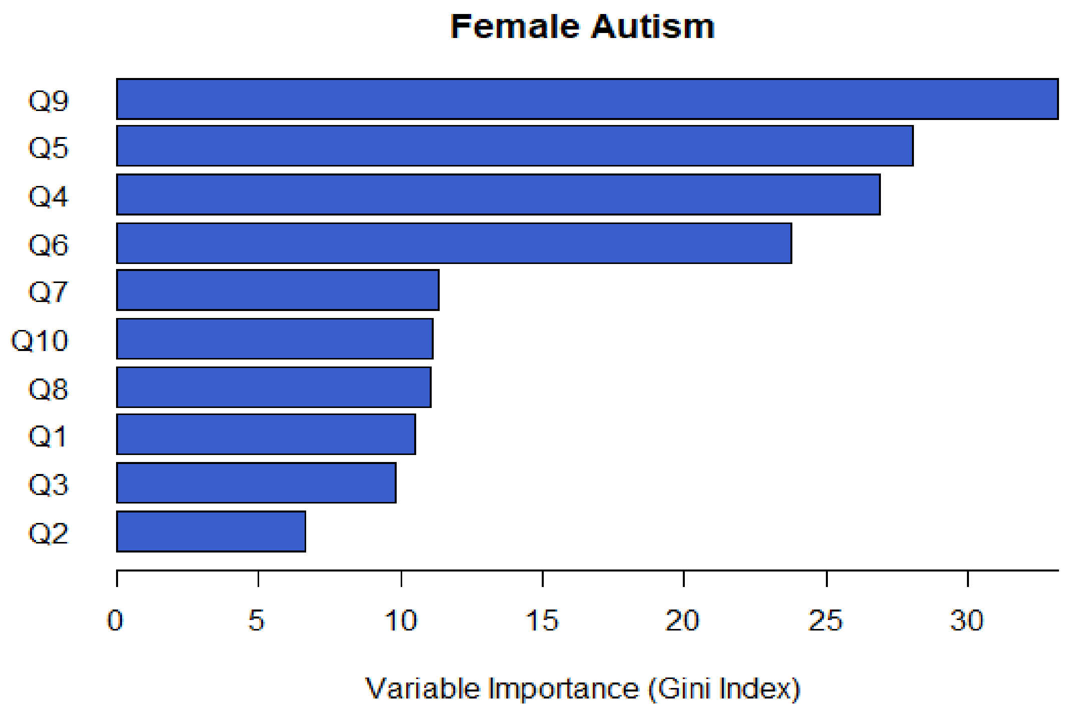
The proportion of correct diagnosis can be calculated as follows:

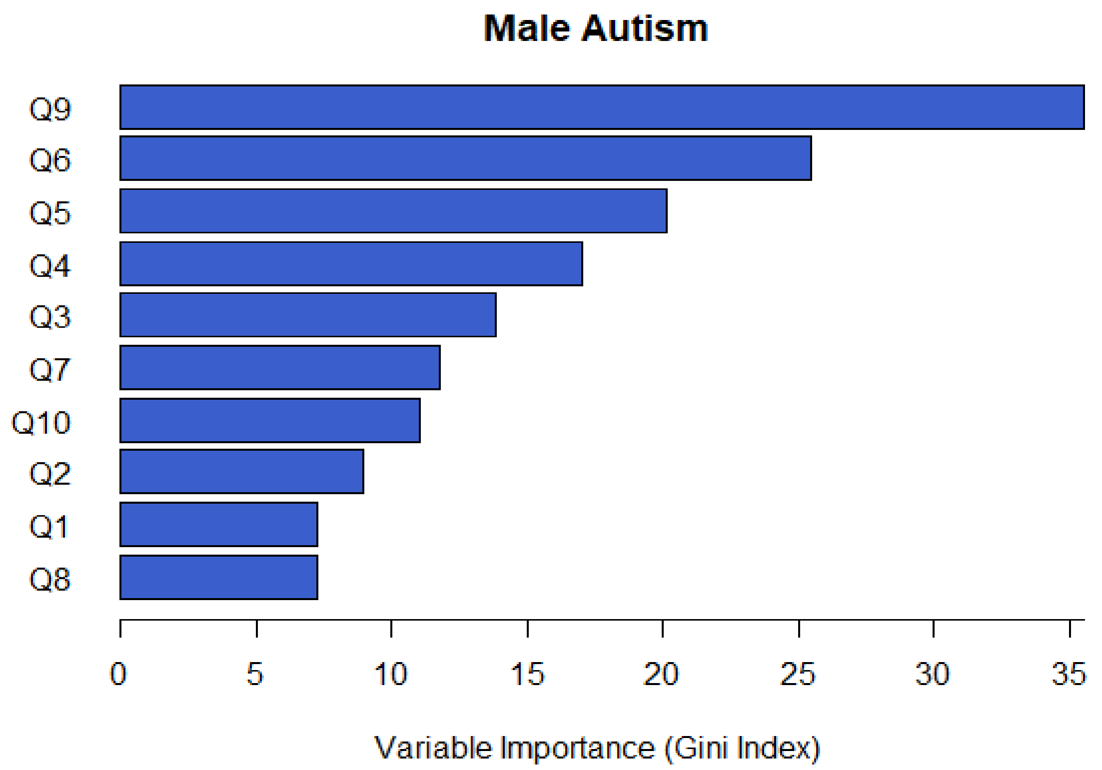
a) Females = (40 + 26) / (40 + 26 + 1 + 4) ≈ 93%

b) Males = (46 + 23) / (46 + 23 + 1 + 1) ≈ 97%

The diagnosis of ASD through the current DSM-5 criteria seems to be quite accurate, as can be seen from the confusion matrix where number of misdiagnosis for both genders are very small in both number and probability. However, the difference of 4% in correct diagnosis probability between males and females, may indicate that the DSM-5 is still slightly more relevant in diagnosing males with ASD than it is for females.

Variance Importance Plots:





The plot above shows importance of variables based on both females and males. As stated in the decision tree, Q9 is the most important question (or variable) in diagnosis of both genders. This characteristic can also be seen in this plot because Q9 has the highest Gini Index.

For females, Q4, Q5, Q6, Q7 and Q10, which are not related to social deficit seems to hold greater significance in diagnosing a female ASD patient (due to the larger Gini index for those questions in females, compared to males). In relation to that, questions, which are social deficit related (Q1, 2, 3, 8) have the lowest Gini Index. This might imply that questions related to social deficit are not so important compared to non-social deficit questions.

The plot for males suggest the similar condition as females. However, social deficit related question, Q3, has relatively larger importance in diagnosing males compared to females.

4. Conclusions and Lesson Learned

Hypothesis 1:

The first hypothesis that females are more likely to be diagnosed as having ASD at a later age compared to males, was validated. This may be because females tend to be misdiagnosed as not having ASD during childhood, where they are less aware of their characteristic symptoms (Howlin & Asgharian, 1999). It may also be because current DSM-5 criteria are more relevant to adult females than younger females, although further research is required to confirm such an inference. Another possible factor may be because females are better at “blending in” and their characteristic symptoms may be hidden from the outside (Lai et al., 2015; Mandy et al., 2012).

The plot will be much more relevant to the hypothesis if proportion is taken in to account. It will also be interesting to reproduce this analysis with a likert scale (ranging from 1 to 7) for each question, so variance in responses can be compared through the ages.

Hypothesis 2:

The second hypothesis was only partially validated. This is because most questions not related to social deficit played a large role in diagnosing *both* females and males with ASD, apart from Q9. It cannot be concretely accepted also because Q9 is still the most important question for diagnosis of both genders. Q9 is related to symptoms regarding social and occupational impairment. Nonetheless, greater weighting was still placed on questions unrelated to social deficit for female diagnosis because those questions were slightly more significant to their diagnosis (larger mean decrease in Gini). It was also the case that questions related to social deficit (Q1, 2, 3, 8) had the lowest importance in females, whereas for males, Q1, 2, 8 but not 3 had the lowest importance. There is thus some reason to suspect that perhaps questions unrelated to social deficit may play a large role in not only diagnosing females, but also males with ASD. If this were indeed the case, further analysis on how social deficit questions (relevant questions except Q9), play a role in diagnosis should be of critical interest.

Furthermore, the large difference in Gini Index of Q6 and Q7 for females should be investigated further to understand the drastic drop in average decrease of Gini index.

Difficulties while obtaining results:

Logistic regression was attempted, but the prediction error seems to be too inaccurate for our results. The range of threshold for the roc plot resulted in an absurd accuracy and shape of the graph.

K-nearest neighbour (KNN) algorithm was attempted as well. However, knowledge on implementation of categorical variables using KNN were lacking. According to research, categorical variables are compatible with KNN, but dummy variables need to be implemented beforehand, for it to work. Therefore, more learning and understanding of dummy variables is required for KNN to work with our dataset.

On top of this, knowledge on functionality of package ggplot2 should be increased to fully apprehend the meaning of our code. This must be done to prevent inefficiency of coding and a better presentation of visualisation.

5. Appendix

```
library(Hmisc)

## Warning: package 'Hmisc' was built under R version 3.4.4

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.4.4

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

library(vcd)

## Warning: package 'vcd' was built under R version 3.4.4

## Loading required package: grid

library(vcdExtra)

## Warning: package 'vcdExtra' was built under R version 3.4.4

## Loading required package: gnm

## Warning: package 'gnm' was built under R version 3.4.4

##
## Attaching package: 'gnm'

## The following object is masked from 'package:lattice':
##
##     barley

library(ggplot2)
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 3.4.4

library(rpart)
library(randomForest)

## Warning: package 'randomForest' was built under R version 3.4.4

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

library(partykit)

## Warning: package 'partykit' was built under R version 3.4.4

## Loading required package: libcoin

## Warning: package 'libcoin' was built under R version 3.4.4

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 3.4.3

library(rattle)

## Warning: package 'rattle' was built under R version 3.4.4

## Rattle: A free graphical interface for data science with R.
## Version 5.1.0 Copyright (c) 2006-2017 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

##
## Attaching package: 'rattle'

## The following object is masked from 'package:randomForest':
##
##     importance

library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 3.4.4

library(RColorBrewer)
library(pROC)

## Warning: package 'pROC' was built under R version 3.4.4

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Understanding Data

```
adolescent <- read.csv("Adolescent_Autism.csv")
adult <- read.csv("Adult_Autism.csv")
child <- read.csv("Child_Autism.csv")

autism <- rbind(adolescent,adult, child)
str(autism)

## 'data.frame':    1100 obs. of  21 variables:
##  $ Q1            : int  0 0 0 0 1 1 0 1 1 0 ...
##  $ Q2            : int  0 0 0 1 1 0 0 1 1 1 ...
##  $ Q3            : int  0 0 0 1 1 0 0 0 1 1 ...
##  $ Q4            : int  1 0 0 1 1 0 1 1 1 0 ...
##  $ Q5            : int  1 0 0 1 1 0 1 1 1 0 ...
##  $ Q6            : int  1 0 0 1 1 1 1 0 1 1 ...
##  $ Q7            : int  1 0 0 0 1 0 1 1 0 0 ...
##  $ Q8            : int  1 0 0 1 0 0 1 1 0 0 ...
##  $ Q9            : int  1 1 1 1 0 1 1 0 0 1 ...
##  $ Q10           : int  0 1 1 0 0 0 0 1 0 0 ...
##  $ Age           : chr  "15" "15" "12" "14" ...
##  $ Gender        : Factor w/ 2 levels "f","m": 2 2 1 1 1 1 1 1 2 1
 ...
##  $ Race          : Factor w/ 12 levels "'Middle Eastern '",..: 6 5 3
 9 3 3 3 1 5 2 ...
##  $ Jaundice      : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 2
1 ...
##  $ FamilyPDD     : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 2
1 ...
##  $ Residence     : Factor w/ 89 levels "'New Zealand'",..: 13 13 9 4
 8 18 7 12 16 14 ...
##  $ SecondUse     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1
1 ...
##  $ ScreeningScore: int  6 2 2 7 7 3 6 7 6 4 ...
##  $ AgeRange      : Factor w/ 4 levels "'12-15 years'",..: 2 2 2 2 2
2 2 2 2 2 ...
##  $ Response      : Factor w/ 7 levels "'Health care professional'
",..: 4 5 2 6 2 2 2 4 4 4 ...
##  $ ASD           : Factor w/ 2 levels "NO","YES": 1 1 1 2 2 1 1 2 1
1 ...

adolescent$lifeStage <- "Adolescent"
adult$lifeStage <- "Adult"
child$lifeStage <- "Child"

autism <- rbind(adolescent,adult, child)
str(autism)

## 'data.frame':    1100 obs. of  22 variables:
##  $ Q1            : int  0 0 0 0 1 1 0 1 1 0 ...
```

```
##  $ Q2           : int  0 0 0 1 1 0 0 1 1 1 ...
##  $ Q3           : int  0 0 0 1 1 0 0 0 1 1 ...
##  $ Q4           : int  1 0 0 1 1 0 1 1 1 0 ...
##  $ Q5           : int  1 0 0 1 1 0 1 1 1 0 ...
##  $ Q6           : int  1 0 0 1 1 1 1 0 1 1 ...
##  $ Q7           : int  1 0 0 0 1 0 1 1 0 0 ...
##  $ Q8           : int  1 0 0 1 0 0 1 1 0 0 ...
##  $ Q9           : int  1 1 1 1 0 1 1 0 0 1 ...
##  $ Q10          : int  0 1 1 0 0 0 0 1 0 0 ...
##  $ Age          : chr  "15" "15" "12" "14" ...
##  $ Gender       : Factor w/ 2 levels "f","m": 2 2 1 1 1 1 1 1 2 1
 ...
##  $ Race         : Factor w/ 12 levels "'Middle Eastern '",..: 6 5 3
 9 3 3 3 1 5 2 ...
##  $ Jaundice     : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 2
1 ...
##  $ FamilyPDD    : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 2
1 ...
##  $ Residence    : Factor w/ 89 levels "'New Zealand'",..: 13 13 9 4
 8 18 7 12 16 14 ...
##  $ SecondUse    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1
1 ...
##  $ ScreeningScore: int  6 2 2 7 7 3 6 7 6 4 ...
##  $ AgeRange     : Factor w/ 4 levels "'12-15 years'",..: 2 2 2 2 2
2 2 2 2 2 ...
##  $ Response     : Factor w/ 7 levels "'Health care professional'
",..: 4 5 2 6 2 2 2 4 4 4 ...
##  $ ASD          : Factor w/ 2 levels "NO","YES": 1 1 1 2 2 1 1 2 1
1 ...
##  $ lifeStage    : chr  "Adolescent" "Adolescent" "Adolescent" "Adol
escent" ...
```

```
dim(autism)
```

```
## [1] 1100   22
```

```
colnames(autism)
```

```
##  [1] "Q1"             "Q2"             "Q3"             "Q4"

##  [5] "Q5"             "Q6"             "Q7"             "Q8"

##  [9] "Q9"             "Q10"            "Age"            "Gender"

## [13] "Race"           "Jaundice"       "FamilyPDD"      "Residence"

## [17] "SecondUse"      "ScreeningScore" "AgeRange"       "Response"

## [21] "ASD"            "lifeStage"
```

```
summary(autism)
```

```
##        Q1               Q2               Q3               Q4
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00
```

```
##  Median :1.0000   Median :0.0000   Median :1.0000   Median :1.00
##  Mean   :0.6991   Mean   :0.4827   Mean   :0.5518   Mean   :0.53
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00
##
##        Q5               Q6               Q7               Q8
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :0.5873   Mean   :0.4436   Mean   :0.4773   Mean   :0.6055
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##        Q9              Q10              Age              Gender
##  Min.   :0.0000   Min.   :0.0000   Length:1100        f:475
##  1st Qu.:0.0000   1st Qu.:0.0000   Class :character   m:625
##  Median :0.0000   Median :1.0000   Mode  :character
##  Mean   :0.4127   Mean   :0.6218
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
##
##                Race     Jaundice  FamilyPDD                 Reside
nce
##  White-European   :381   no :935   no :946   'United States'       :
167
##  Asian            :185   yes:165   yes:154   'United Kingdom'      :
155
##  ?                :144                       India                 :
130
##  'Middle Eastern ':128                       'New Zealand'         :
 95
##  Black            : 65                       'United Arab Emirates':
 90
##  'South Asian'    : 60                       Jordan                :
 68
##  (Other)          :137                       (Other)               :
395
##  SecondUse  ScreeningScore           AgeRange
##  no :1073   Min.   : 0.000   '12-15 years':  7
##  yes:  27   1st Qu.: 3.000   '12-16 years': 97
##             Median : 5.000   '18 and more':704
##             Mean   : 5.412   '4-11 years' :292
##             3rd Qu.: 7.250
##             Max.   :10.000
##
##                          Response    ASD       lifeStage
##  'Health care professional': 23   NO :707   Length:1100
##  ?                         :144   YES:393   Class :character
##  Others                    :  8             Mode  :character
##  Parent                    :300
##  Relative                  : 53
##  Self                      :571
##  self                      :  1
```

```
autism$Q1[autism$Q1==0] <- "no"
autism$Q1[autism$Q1==1] <- "yes"
autism$Q2[autism$Q2==0] <- "no"
autism$Q2[autism$Q2==1] <- "yes"
autism$Q3[autism$Q3==0] <- "no"
autism$Q3[autism$Q3==1] <- "yes"
autism$Q4[autism$Q4==0] <- "no"
autism$Q4[autism$Q4==1] <- "yes"
autism$Q5[autism$Q5==0] <- "no"
autism$Q5[autism$Q5==1] <- "yes"
autism$Q6[autism$Q6==0] <- "no"
autism$Q6[autism$Q6==1] <- "yes"
autism$Q7[autism$Q7==0] <- "no"
autism$Q7[autism$Q7==1] <- "yes"
autism$Q8[autism$Q8==0] <- "no"
autism$Q8[autism$Q8==1] <- "yes"
autism$Q9[autism$Q9==0] <- "no"
autism$Q9[autism$Q9==1] <- "yes"
autism$Q10[autism$Q10==0] <- "no"
autism$Q10[autism$Q10==1] <- "yes"

autism$Q1 <- as.factor(autism$Q1)
autism$Q2 <- as.factor(autism$Q2)
autism$Q3 <- as.factor(autism$Q3)
autism$Q4 <- as.factor(autism$Q4)
autism$Q5 <- as.factor(autism$Q5)
autism$Q6 <- as.factor(autism$Q6)
autism$Q7 <- as.factor(autism$Q7)
autism$Q8 <- as.factor(autism$Q8)
autism$Q9 <- as.factor(autism$Q9)
autism$Q10 <- as.factor(autism$Q10)

female <- subset(autism, subset = Gender=="f")
male <- subset(autism, subset = Gender =="m")

describe(female)

## female
##
##  22  Variables      475  Observations
## --------------------------------------------------------------------
-------
## Q1
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    122   353
## Proportion 0.257 0.743
## --------------------------------------------------------------------
-------
## Q2
##        n  missing distinct
##      475        0        2
##
```

```
## Value          no   yes
## Frequency    242   233
## Proportion 0.509 0.491
## -----------------------------------------------------------------------
-------
## Q3
##        n  missing distinct
##      475        0        2
##
## Value          no   yes
## Frequency    227   248
## Proportion 0.478 0.522
## -----------------------------------------------------------------------
-------
## Q4
##        n  missing distinct
##      475        0        2
##
## Value          no   yes
## Frequency    213   262
## Proportion 0.448 0.552
## -----------------------------------------------------------------------
-------
## Q5
##        n  missing distinct
##      475        0        2
##
## Value          no   yes
## Frequency    199   276
## Proportion 0.419 0.581
## -----------------------------------------------------------------------
-------
## Q6
##        n  missing distinct
##      475        0        2
##
## Value         no  yes
## Frequency   266  209
## Proportion 0.56 0.44
## -----------------------------------------------------------------------
-------
## Q7
##        n  missing distinct
##      475        0        2
##
## Value          no   yes
## Frequency    272   203
## Proportion 0.573 0.427
## -----------------------------------------------------------------------
-------
## Q8
##        n  missing distinct
##      475        0        2
##
```

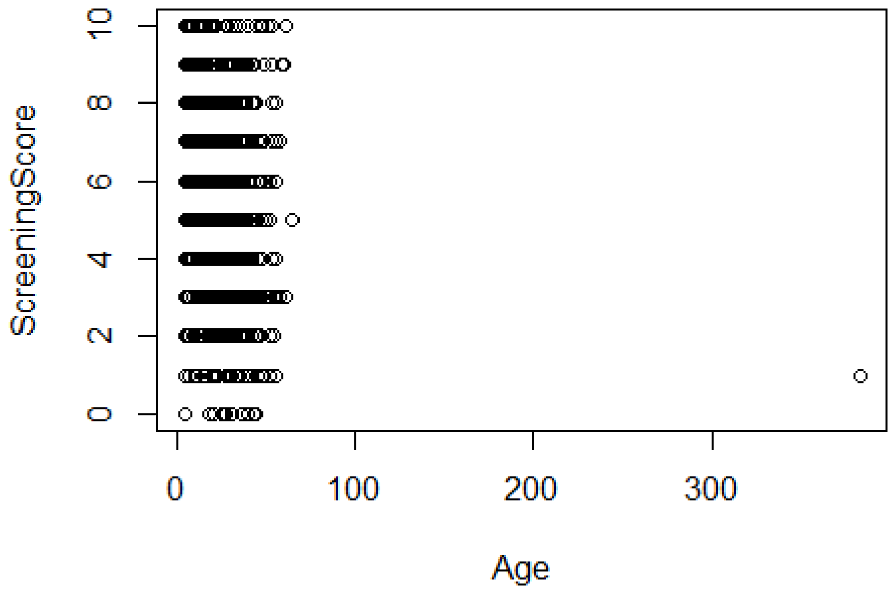
```
## Value         no   yes
## Frequency    198   277
## Proportion 0.417 0.583
## -----------------------------------------------------------------------
-------
## Q9
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    275   200
## Proportion 0.579 0.421
## -----------------------------------------------------------------------
-------
## Q10
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    174   301
## Proportion 0.366 0.634
## -----------------------------------------------------------------------
-------
## Age
##        n  missing distinct
##      475        0       55
##
## lowest : ?  10 11 12 13, highest: 60 61 7  8  9
## -----------------------------------------------------------------------
-------
## Gender
##        n  missing distinct    value
##      475        0        1        f
##
## Value        f
## Frequency  475
## Proportion   1
## -----------------------------------------------------------------------
-------
## Race
##        n  missing distinct
##      475        0       11
##
## 'Middle Eastern ' (54, 0.114), 'South Asian' (26, 0.055), ? (64, 0.1
35),
## Asian (61, 0.128), Black (32, 0.067), Hispanic (5, 0.011), Latino (1
2,
## 0.025), Others (28, 0.059), White-European (185, 0.389), Pasifika
(6,
## 0.013), Turkish (2, 0.004)
## -----------------------------------------------------------------------
-------
## Jaundice
##        n  missing distinct
```

```
##      475        0        2
##
## Value         no   yes
## Frequency    408    67
## Proportion 0.859 0.141
## --------------------------------------------------------------------------
-------
## FamilyPDD
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    394    81
## Proportion 0.829 0.171
## --------------------------------------------------------------------------
-------
## Residence
##        n  missing distinct
##      475        0       60
##
## lowest : 'New Zealand'          'South Africa'         'United Arab
Emirates' 'United Kingdom'       'United States'
## highest: Bhutan                 Georgia                Kuwait
          Nigeria               Syria
## --------------------------------------------------------------------------
-------
## SecondUse
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    466     9
## Proportion 0.981 0.019
## --------------------------------------------------------------------------
-------
## ScreeningScore
##        n  missing distinct     Info     Mean      Gmd      .05
 .10
##      475        0       11    0.987    5.394    2.923      1.7
2.0
##      .25      .50      .75      .90      .95
##      3.0      5.0      8.0      9.0      9.0
##

## Value          0     1     2     3     4     5     6     7     8
 9
## Frequency      8    16    46    51    80    48    49    55    57
44
## Proportion 0.017 0.034 0.097 0.107 0.168 0.101 0.103 0.116 0.120 0.0
93
##
## Value         10
## Frequency     21
## Proportion 0.044
```

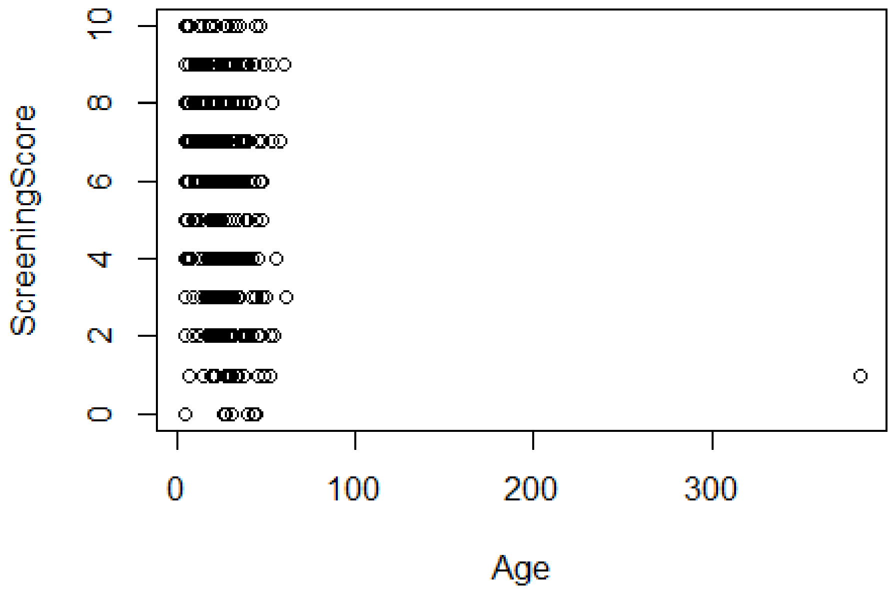
```
## ---------------------------------------------------------------------
-------
## AgeRange
##        n  missing distinct
##      475        0        4
##
## Value      '12-15 years' '12-16 years' '18 and more'  '4-11 years'
## Frequency              5            49           337            84
## Proportion         0.011         0.103         0.709         0.177
## ---------------------------------------------------------------------
-------
## Response
##        n  missing distinct
##      475        0        5
##
## 'Health care professional' (10, 0.021), ? (64, 0.135), Parent (105,
## 0.221), Relative (12, 0.025), Self (284, 0.598)
## ---------------------------------------------------------------------
-------
## ASD
##        n  missing distinct
##      475        0        2
##
## Value         NO   YES
## Frequency    298   177
## Proportion 0.627 0.373
## ---------------------------------------------------------------------
-------
## lifeStage
##        n  missing distinct
##      475        0        3
##
## Value      Adolescent      Adult      Child
## Frequency          54        337         84
## Proportion      0.114      0.709      0.177
## ---------------------------------------------------------------------
-------
```

Data Wrangling / Data Cleaning

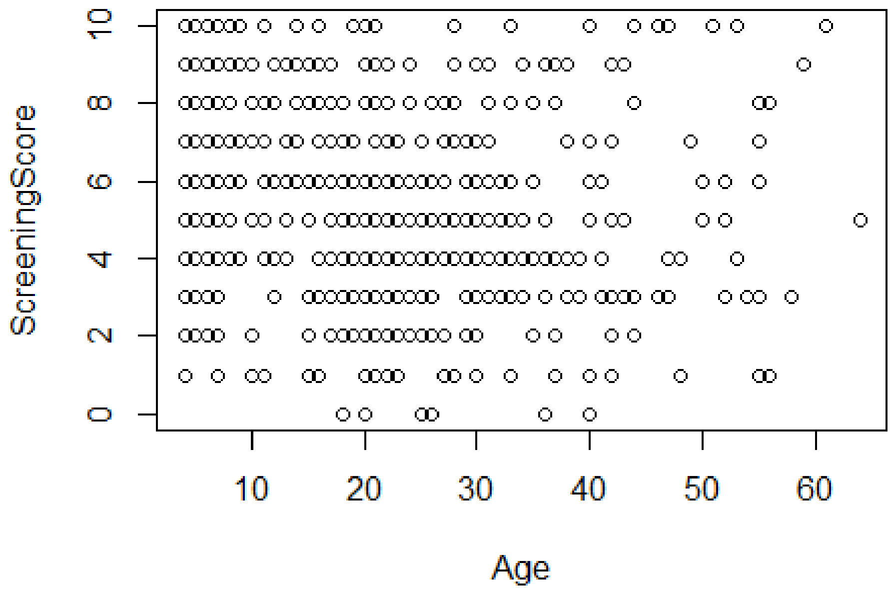
```
plot(ScreeningScore~Age, data = autism)

## Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by c
oercion
```




```
autism <- autism[-157, ]
autism <- autism[-947, ]
dim(autism)

## [1] 1098   22

plot(ScreeningScore~Age, data = female)

## Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by c
oercion
```




```r
plot(ScreeningScore~Age, data = male)

## Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by c
oercion
```




```r
male <- male[-52:-79, ]

male <- male[-419:-540, ]
```

```r
dim(male)

## [1] 475  22

describe(male)

## male
##
##  22  Variables      475  Observations
## ---------------------------------------------------------------------
-------
## Q1
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    145   330
## Proportion 0.305 0.695
## ---------------------------------------------------------------------
-------
## Q2
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    251   224
## Proportion 0.528 0.472
## ---------------------------------------------------------------------
-------
## Q3
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    224   251
## Proportion 0.472 0.528
## ---------------------------------------------------------------------
-------
## Q4
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    236   239
## Proportion 0.497 0.503
## ---------------------------------------------------------------------
-------
## Q5
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    211   264
## Proportion 0.444 0.556
```

```
## ---------------------------------------------------------------------
-------
## Q6
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    290   185
## Proportion 0.611 0.389
## ---------------------------------------------------------------------
-------
## Q7
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    236   239
## Proportion 0.497 0.503
## ---------------------------------------------------------------------
-------
## Q8
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    170   305
## Proportion 0.358 0.642
## ---------------------------------------------------------------------
-------
## Q9
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    293   182
## Proportion 0.617 0.383
## ---------------------------------------------------------------------
-------
## Q10
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    199   276
## Proportion 0.419 0.581
## ---------------------------------------------------------------------
-------
## Age
##        n  missing distinct
##      475        0       56
##
## lowest : ?  10 11 12 13, highest: 6  61 7  8  9
## ---------------------------------------------------------------------
-------
```

```
## Gender
##        n  missing distinct    value
##      475        0        1        m
##
## Value        m
## Frequency  475
## Proportion   1
## ----------------------------------------------------------------------
-------
## Race
##        n  missing distinct
##      475        0       12
##
## 'Middle Eastern ' (61, 0.128), 'South Asian' (27, 0.057), ? (58, 0.1
22),
## Asian (98, 0.206), Black (27, 0.057), Hispanic (15, 0.032), Latino
(14,
## 0.029), Others (24, 0.051), White-European (140, 0.295), others (1,
## 0.002), Pasifika (5, 0.011), Turkish (5, 0.011)
## ----------------------------------------------------------------------
-------
## Jaundice
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    409    66
## Proportion 0.861 0.139
## ----------------------------------------------------------------------
-------
## FamilyPDD
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
## Frequency    428    47
## Proportion 0.901 0.099
## ----------------------------------------------------------------------
-------
## Residence
##        n  missing distinct
##      475        0       62
##
## lowest : 'New Zealand'          'South Africa'         'United Arab
Emirates' 'United Kingdom'       'United States'
## highest: Bulgaria               Europe                 Ghana
          Libya                  Malta
## ----------------------------------------------------------------------
-------
## SecondUse
##        n  missing distinct
##      475        0        2
##
## Value         no   yes
```

```
## Frequency    464    11
## Proportion 0.977 0.023
## ---------------------------------------------------------------------
-------
## ScreeningScore
##        n  missing distinct     Info     Mean      Gmd      .05
 .10
##      475        0       11    0.985    5.253    2.828        1
  2
##      .25      .50      .75      .90      .95
##        3        5        7        9       10
##

## Value          0     1     2     3     4     5     6     7     8
 9
## Frequency      5    20    32    75    76    63    56    40    46
37
## Proportion 0.011 0.042 0.067 0.158 0.160 0.133 0.118 0.084 0.097 0.0
78
##
## Value         10
## Frequency     25
## Proportion 0.053
## ---------------------------------------------------------------------
-------
## AgeRange
##        n  missing distinct
##      475        0        4
##
## Value      '12-15 years' '12-16 years' '18 and more'  '4-11 years'
## Frequency             2            48           339            86
## Proportion         0.004         0.101         0.714         0.181
## ---------------------------------------------------------------------
-------
## Response
##        n  missing distinct
##      475        0        6
##
## 'Health care professional' (9, 0.019), ? (58, 0.122), Others (8, 0.0
17),
## Parent (99, 0.208), Relative (27, 0.057), Self (274, 0.577)
## ---------------------------------------------------------------------
-------
## ASD
##        n  missing distinct
##      475        0        2
##
## Value         NO   YES
## Frequency    327   148
## Proportion 0.688 0.312
## ---------------------------------------------------------------------
-------
## lifeStage
##        n  missing distinct
```

```
##      475        0        3
##
## Value      Adolescent      Adult      Child
## Frequency          50        339         86
## Proportion      0.105      0.714      0.181
## ----------------------------------------------------------------------
-------
```

```
autism <- rbind(male, female)
str(autism)
```

```
## 'data.frame':    950 obs. of  22 variables:
##  $ Q1            : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 2 2 2
1 ...
##  $ Q2            : Factor w/ 2 levels "no","yes": 1 1 2 2 2 1 2 2 2
2 ...
##  $ Q3            : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 2 1 1
2 ...
##  $ Q4            : Factor w/ 2 levels "no","yes": 2 1 2 1 2 2 2 2 1
2 ...
##  $ Q5            : Factor w/ 2 levels "no","yes": 2 1 2 2 1 2 2 1 1
2 ...
##  $ Q6            : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 1 1
2 ...
##  $ Q7            : Factor w/ 2 levels "no","yes": 2 1 1 2 2 2 2 1 2
2 ...
##  $ Q8            : Factor w/ 2 levels "no","yes": 2 1 1 1 2 2 2 1 1
2 ...
##  $ Q9            : Factor w/ 2 levels "no","yes": 2 2 1 2 2 2 2 1 2
2 ...
##  $ Q10           : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 2 1 1
2 ...
##  $ Age           : chr  "15" "15" "12" "12" ...
##  $ Gender        : Factor w/ 2 levels "f","m": 2 2 2 2 2 2 2 2 2 2
 ...
##  $ Race          : Factor w/ 12 levels "'Middle Eastern '",..: 6 5 5
 9 9 1 6 1 3 8 ...
##  $ Jaundice      : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 1 1
1 ...
##  $ FamilyPDD     : Factor w/ 2 levels "no","yes": 2 1 2 1 1 2 1 1 1
1 ...
##  $ Residence     : Factor w/ 89 levels "'New Zealand'",..: 13 13 16
4 1 5 11 30 7 4 ...
##  $ SecondUse     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1
1 ...
##  $ ScreeningScore: int  6 2 6 8 9 9 10 3 4 9 ...
##  $ AgeRange      : Factor w/ 4 levels "'12-15 years'",..: 2 2 2 2 2
2 2 2 2 2 ...
##  $ Response      : Factor w/ 7 levels "'Health care professional'
",..: 4 5 4 4 4 4 6 6 2 4 ...
##  $ ASD           : Factor w/ 2 levels "NO","YES": 1 1 1 2 2 2 2 1 1
2 ...
##  $ lifeStage     : chr  "Adolescent" "Adolescent" "Adolescent" "Adol
escent" ...
```

```
des <- describe(autism)
des

## autism
##
##  22  Variables      950  Observations
## ------------------------------------------------------------------------
-------
## Q1
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    267   683
## Proportion 0.281 0.719
## ------------------------------------------------------------------------
-------
## Q2
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    493   457
## Proportion 0.519 0.481
## ------------------------------------------------------------------------
-------
## Q3
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    451   499
## Proportion 0.475 0.525
## ------------------------------------------------------------------------
-------
## Q4
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    449   501
## Proportion 0.473 0.527
## ------------------------------------------------------------------------
-------
## Q5
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    410   540
## Proportion 0.432 0.568
## ------------------------------------------------------------------------
-------
## Q6
```

```
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    556   394
## Proportion 0.585 0.415
## ----------------------------------------------------------------------
-------
## Q7
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    508   442
## Proportion 0.535 0.465
## ----------------------------------------------------------------------
-------
## Q8
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    368   582
## Proportion 0.387 0.613
## ----------------------------------------------------------------------
-------
## Q9
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    568   382
## Proportion 0.598 0.402
## ----------------------------------------------------------------------
-------
## Q10
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    373   577
## Proportion 0.393 0.607
## ----------------------------------------------------------------------
-------
## Age
##        n  missing distinct
##      950        0       59
##
## lowest : ?  10 11 12 13, highest: 60 61 7  8  9
## ----------------------------------------------------------------------
-------
## Gender
##        n  missing distinct
##      950        0        2
```

```
##
## Value        f   m
## Frequency  475 475
## Proportion 0.5 0.5
## ----------------------------------------------------------------------
-------
## Race
##        n  missing distinct
##      950        0       12
##
## 'Middle Eastern ' (115, 0.121), 'South Asian' (53, 0.056), ? (122,
0.128),
## Asian (159, 0.167), Black (59, 0.062), Hispanic (20, 0.021), Latino
(26,
## 0.027), Others (52, 0.055), White-European (325, 0.342), others (1,
## 0.001), Pasifika (11, 0.012), Turkish (7, 0.007)
## ----------------------------------------------------------------------
-------
## Jaundice
##        n  missing distinct
##      950        0        2
##
## Value        no  yes
## Frequency   817  133
## Proportion 0.86 0.14
## ----------------------------------------------------------------------
-------
## FamilyPDD
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    822   128
## Proportion 0.865 0.135
## ----------------------------------------------------------------------
-------
## Residence
##        n  missing distinct
##      950        0       84
##
## lowest : 'New Zealand'          'South Africa'         'United Arab
Emirates' 'United Kingdom'       'United States'
## highest: Kuwait                 Libya                  Malta
          Nigeria                Syria
## ----------------------------------------------------------------------
-------
## SecondUse
##        n  missing distinct
##      950        0        2
##
## Value         no   yes
## Frequency    930    20
## Proportion 0.979 0.021
## ----------------------------------------------------------------------
```

```
-------
## ScreeningScore
##        n  missing distinct     Info     Mean      Gmd      .05
 .10
##      950        0       11    0.987    5.323    2.877        1
  2
##      .25      .50      .75      .90      .95
##        3        5        7        9        9
##

## Value          0     1     2     3     4     5     6     7     8
 9
## Frequency     13    36    78   126   156   111   105    95   103
81
## Proportion 0.014 0.038 0.082 0.133 0.164 0.117 0.111 0.100 0.108 0.0
85
##
## Value         10
## Frequency     46
## Proportion 0.048
## ----------------------------------------------------------------------
-------
## AgeRange
##        n  missing distinct
##      950        0        4
##
## Value      '12-15 years' '12-16 years' '18 and more'  '4-11 years'
## Frequency              7            97           676           170
## Proportion         0.007         0.102         0.712         0.179
## ----------------------------------------------------------------------
-------
## Response
##        n  missing distinct
##      950        0        6
##
## 'Health care professional' (19, 0.020), ? (122, 0.128), Others (8,
0.008),
## Parent (204, 0.215), Relative (39, 0.041), Self (558, 0.587)
## ----------------------------------------------------------------------
-------
## ASD
##        n  missing distinct
##      950        0        2
##
## Value         NO   YES
## Frequency    625   325
## Proportion 0.658 0.342
## ----------------------------------------------------------------------
-------
## lifeStage
##        n  missing distinct
##      950        0        3
##
## Value      Adolescent      Adult      Child
```

```
## Frequency          104        676        170
## Proportion       0.109      0.712      0.179
## -----------------------------------------------------------------------
-------
```

```
autism$Gender <- as.character(autism$Gender)
class(autism$Gender)
```

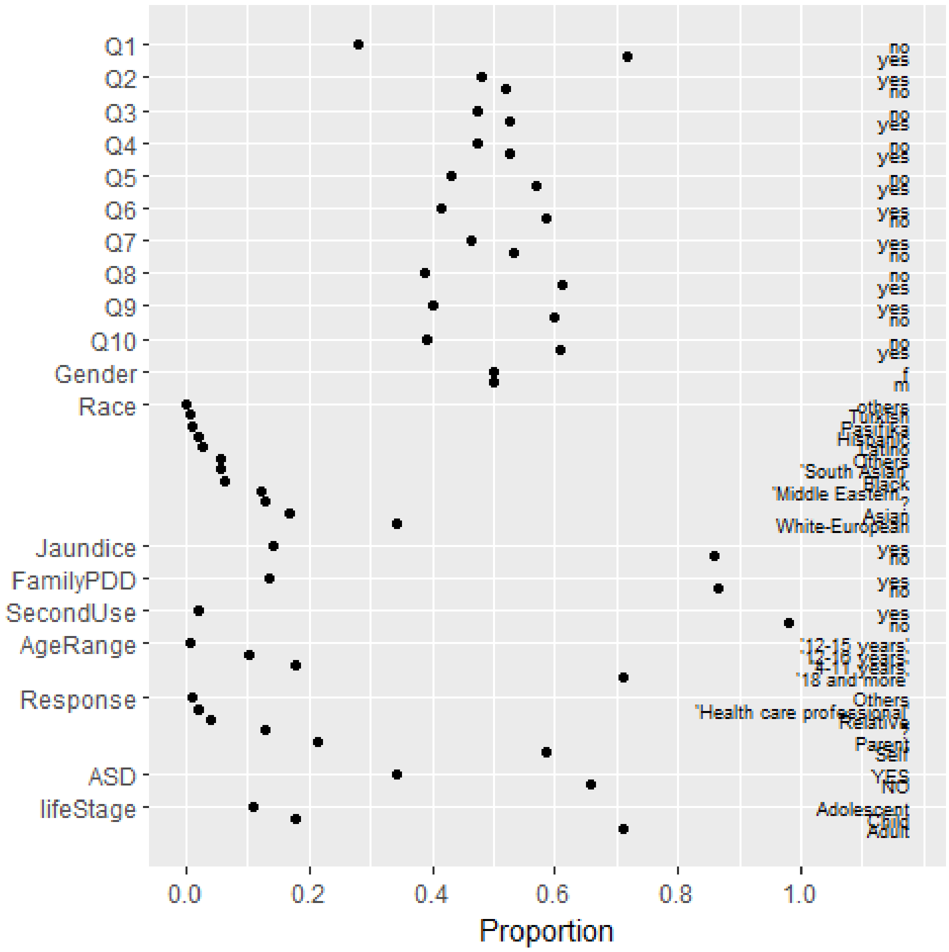
```
## [1] "character"
```

```
str(autism)
```

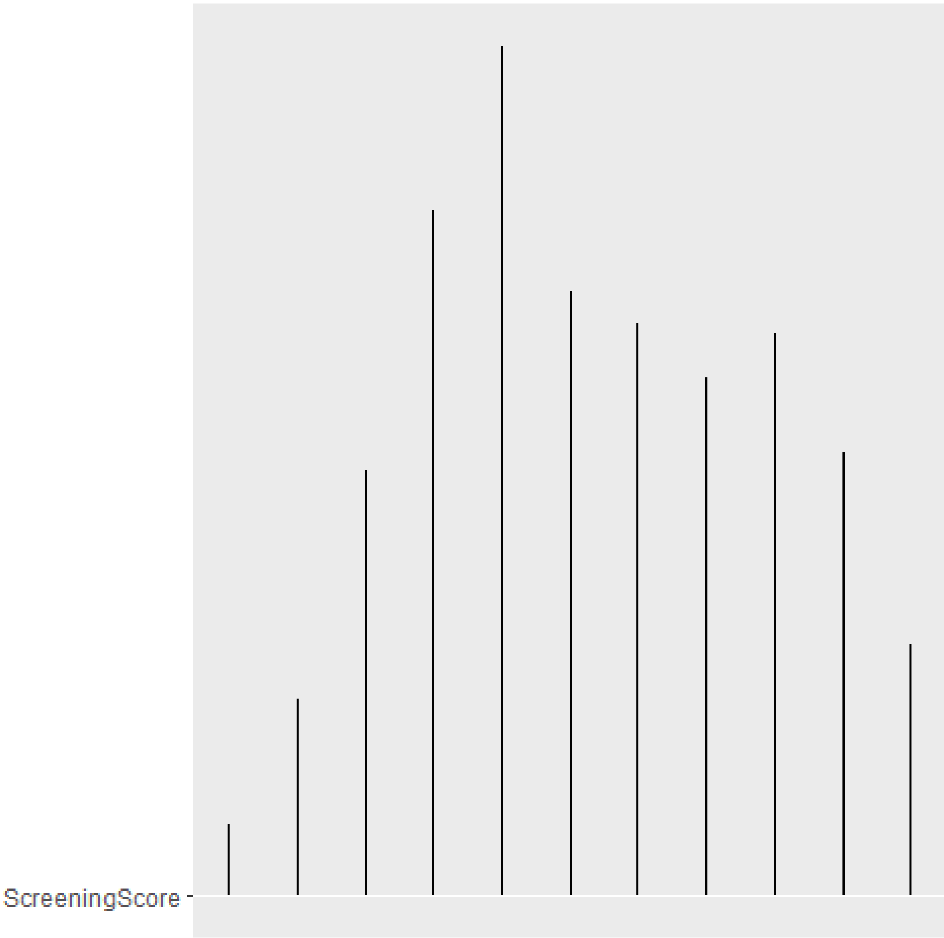
```
## 'data.frame':    950 obs. of  22 variables:
##  $ Q1            : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 2 2 2
1 ...
##  $ Q2            : Factor w/ 2 levels "no","yes": 1 1 2 2 2 1 2 2 2
2 ...
##  $ Q3            : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 2 1 1
2 ...
##  $ Q4            : Factor w/ 2 levels "no","yes": 2 1 2 1 2 2 2 2 1
2 ...
##  $ Q5            : Factor w/ 2 levels "no","yes": 2 1 2 2 1 2 2 1 1
2 ...
##  $ Q6            : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 1 1
2 ...
##  $ Q7            : Factor w/ 2 levels "no","yes": 2 1 1 2 2 2 2 1 2
2 ...
##  $ Q8            : Factor w/ 2 levels "no","yes": 2 1 1 1 2 2 2 1 1
2 ...
##  $ Q9            : Factor w/ 2 levels "no","yes": 2 2 1 2 2 2 2 1 2
2 ...
##  $ Q10           : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 2 1 1
2 ...
##  $ Age           : chr  "15" "15" "12" "12" ...
##  $ Gender        : chr  "m" "m" "m" "m" ...
##  $ Race          : Factor w/ 12 levels "'Middle Eastern '",..: 6 5 5
 9 9 1 6 1 3 8 ...
##  $ Jaundice      : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 1 1
1 ...
##  $ FamilyPDD     : Factor w/ 2 levels "no","yes": 2 1 2 1 1 2 1 1 1
1 ...
##  $ Residence     : Factor w/ 89 levels "'New Zealand'",..: 13 13 16
4 1 5 11 30 7 4 ...
##  $ SecondUse     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1
1 ...
##  $ ScreeningScore: int  6 2 6 8 9 9 10 3 4 9 ...
##  $ AgeRange      : Factor w/ 4 levels "'12-15 years'",..: 2 2 2 2 2
2 2 2 2 2 ...
##  $ Response      : Factor w/ 7 levels "'Health care professional'
",..: 4 5 4 4 4 4 6 6 2 4 ...
##  $ ASD           : Factor w/ 2 levels "NO","YES": 1 1 1 2 2 2 2 1 1
2 ...
##  $ lifeStage     : chr  "Adolescent" "Adolescent" "Adolescent" "Adol
escent" ...
```

EDA

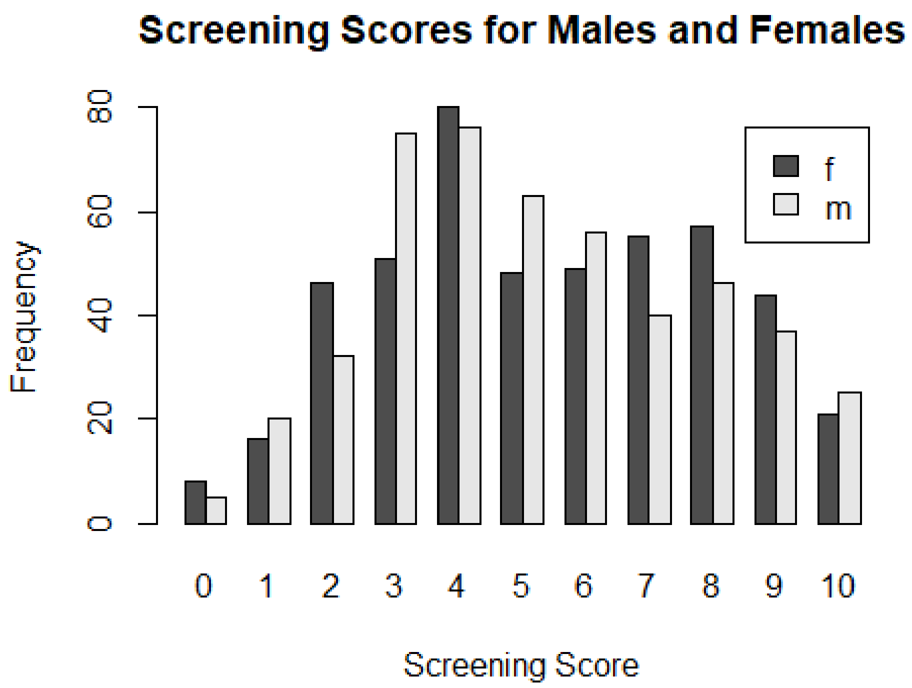
```
plot(des)

## $Categorical
```



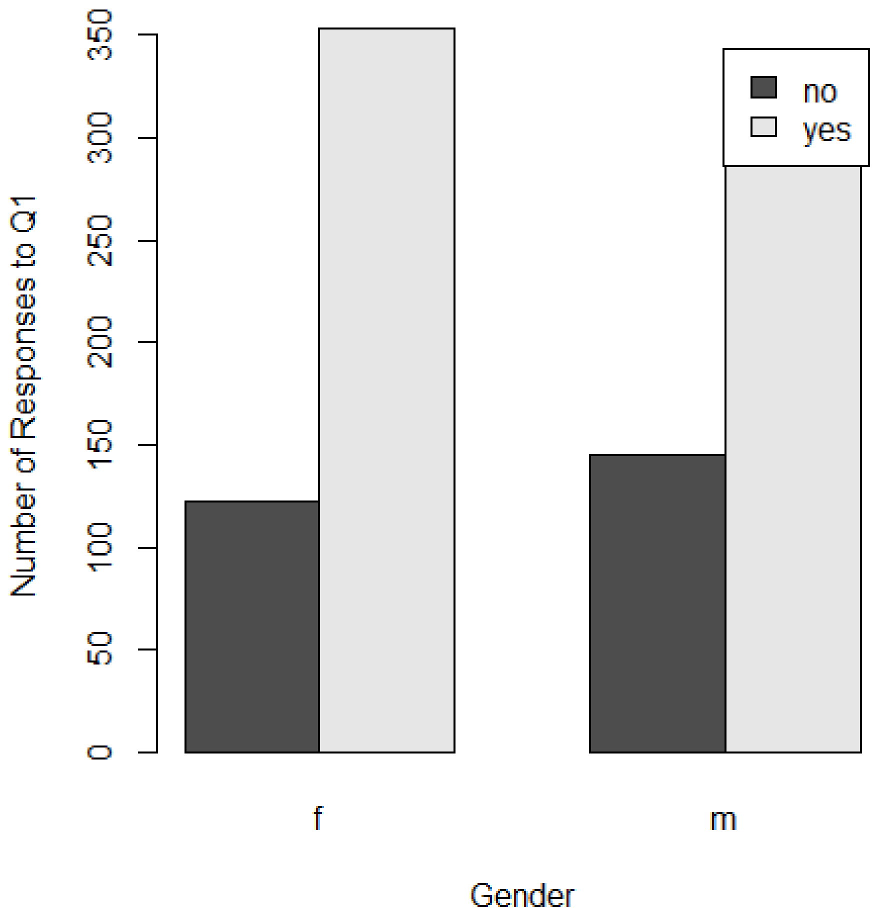

```
##
## $Continuous
```



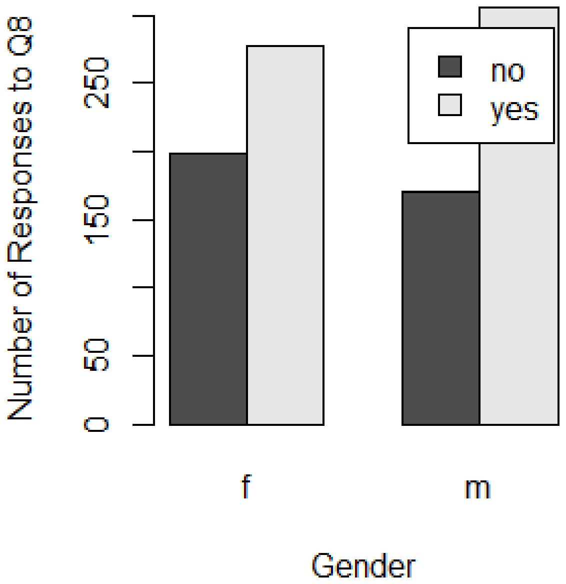

```
table(autism$Gender, autism$ScreeningScore)

##
##       0  1  2  3  4  5  6  7  8  9 10
##    f  8 16 46 51 80 48 49 55 57 44 21
##    m  5 20 32 75 76 63 56 40 46 37 25

barplot(table(autism$Gender, autism$ScreeningScore),beside=TRUE,legend.
text = TRUE,xlab = "Screening Score", ylab = "Frequency",main = "Screen
ing Scores for Males and Females")
```



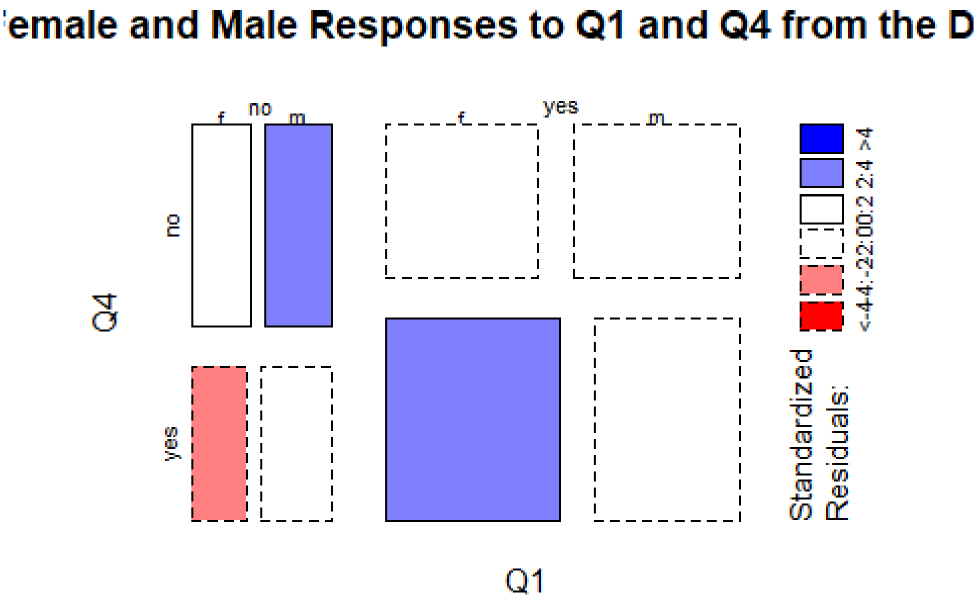

```
par(pty="s")
barplot(table(autism$Q1,autism$Gender), beside = TRUE, legend.text = TRUE, ylab = "Number of Responses to Q1", xlab = "Gender")
```



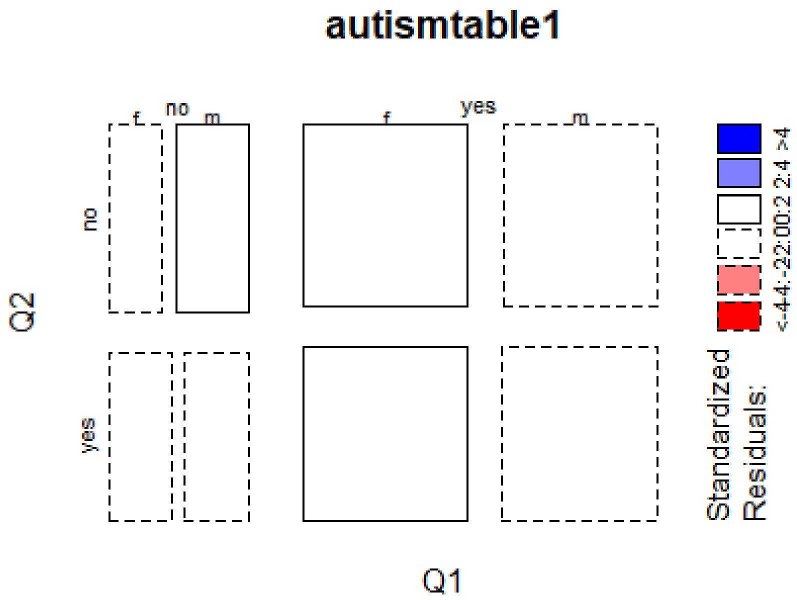

```r
par(pty="s")
barplot(table(autism$Q8,autism$Gender), beside = TRUE, legend.text = TRUE, ylab = "Number of Responses to Q8", xlab = "Gender")
```




```
autismtable <- table(autism$Q1, autism$Q4, autism$Gender, dnn = c("Q1",
 "Q4", "Gender") )
autismtable

## , , Gender = f
##
##      Q4
## Q1     no yes
##   no   71  51
##   yes 142 211
##
## , , Gender = m
##
##      Q4
## Q1     no yes
##   no   81  64
##   yes 155 175

mosaicplot(autismtable, shade = TRUE, main = "Female and Male Responses
 to Q1 and Q4 from the DSM-5")
```



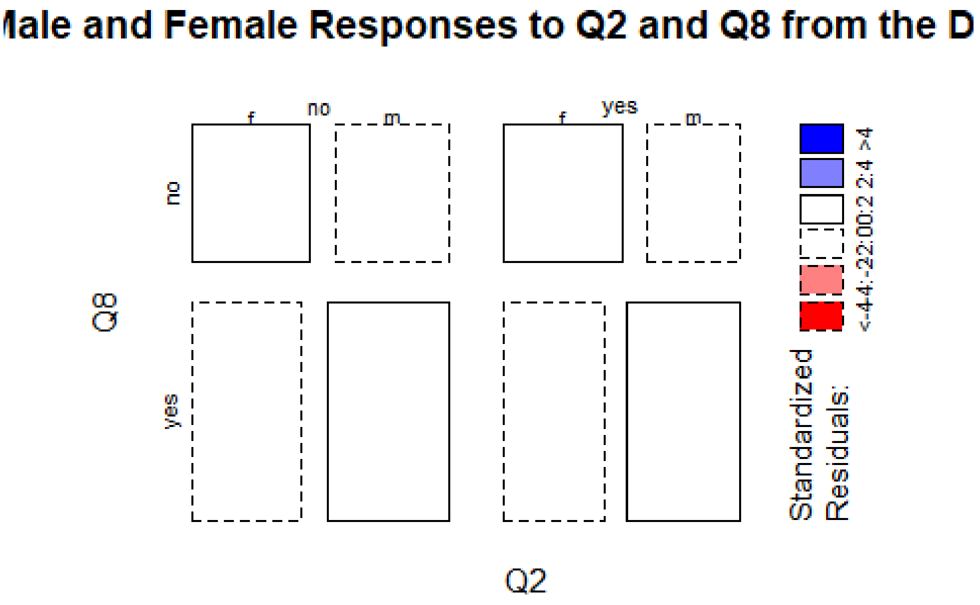

```
autismtable1 <- table(autism$Q1, autism$Q2, autism$Gender, dnn = c("Q1
", "Q2", "Gender") )
autismtable1

## , , Gender = f
##
##      Q2
## Q1     no yes
##   no   60  62
##   yes 182 171
##
## , , Gender = m
##
##      Q2
## Q1     no yes
##   no   82  63
##   yes 169 161

mosaicplot(autismtable1, shade = TRUE)
```




```
autismtable2 <- table(autism$Q2, autism$Q8, autism$Gender, dnn = c("Q2
", "Q8", "Gender") )
autismtable2
```

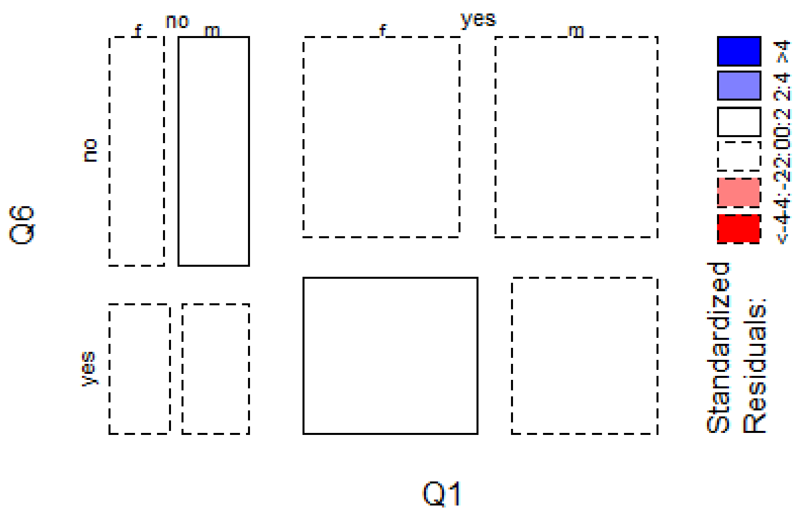
```
## , , Gender = f
##
##      Q8
## Q2     no yes
##   no   98 144
##   yes 100 133
##
## , , Gender = m
##
##      Q8
## Q2     no yes
##   no   93 158
##   yes  77 147
```

```
mosaicplot(autismtable2, shade = TRUE, main = "Male and Female Response
s to Q2 and Q8 from the DSM-5")
```




```
autismtable3 <- table(autism$Q1, autism$Q6, autism$Gender, dnn = c("Q1
", "Q6", "Gender") )
autismtable3

## , , Gender = f
##
##       Q6
## Q1      no yes
##   no    76  46
##   yes 190 163
##
## , , Gender = m
##
##       Q6
## Q1      no yes
##   no    95  50
##   yes 195 135

mosaicplot(autismtable3, shade = TRUE, main = "Male and Female Response
s to Q1 and Q6 from the DSM-5")
```

ale and Female Responses to Q1 and Q6 from the D




```
QFemale <- female[, 1:10]

QMale <- male[, 1:10]

table(QFemale$Q1)

##
##  no yes
## 122 353

table(QMale$Q1)

##
##  no yes
## 145 330

table(QFemale$Q2)

##
##  no yes
## 242 233

table(QMale$Q2)

##
##  no yes
## 251 224

table(QFemale$Q3)

##
##  no yes
## 227 248
```

```r
table(QMale$Q3)

##
##  no yes
## 224 251

table(QFemale$Q4)

##
##  no yes
## 213 262

table(QMale$Q4)

##
##  no yes
## 236 239

table(QFemale$Q5)

##
##  no yes
## 199 276

table(QMale$Q5)

##
##  no yes
## 211 264

table(QFemale$Q6)

##
##  no yes
## 266 209

table(QMale$Q6)

##
##  no yes
## 290 185

table(QFemale$Q7)

##
##  no yes
## 272 203

table(QMale$Q7)

##
##  no yes
## 236 239

table(QFemale$Q8)
```

```
##
##  no yes
## 198 277
```

```
table(QMale$Q8)
```

```
##
##  no yes
## 170 305
```

```
table(QFemale$Q9)
```

```
##
##  no yes
## 275 200
```

```
table(QMale$Q9)
```

```
##
##  no yes
## 293 182
```

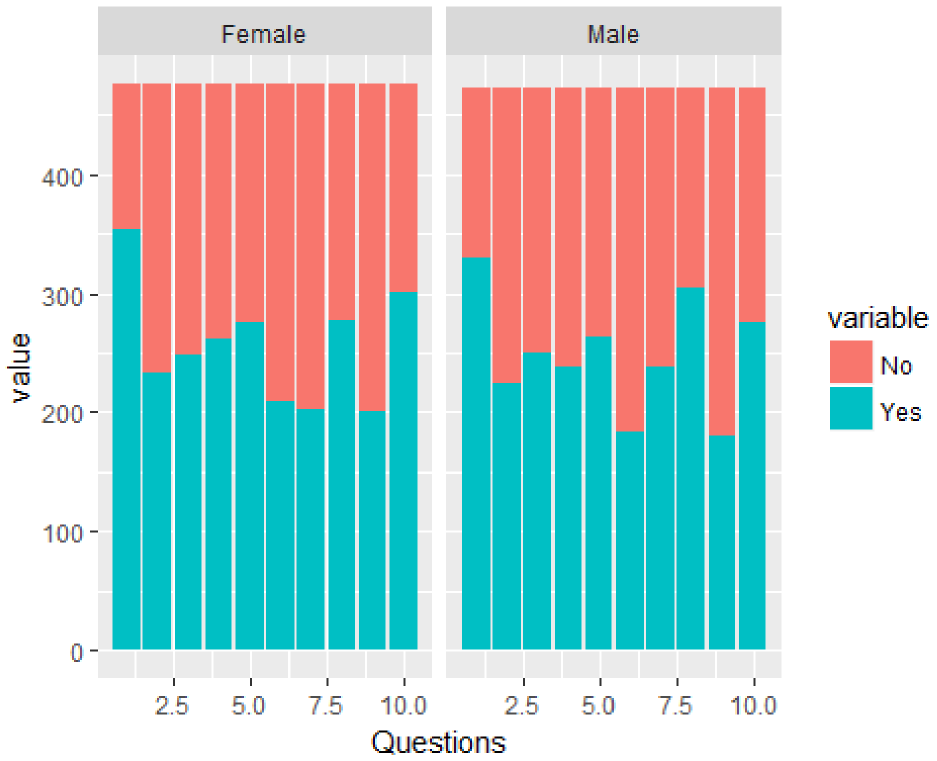
```
table(QFemale$Q10)
```

```
##
##  no yes
## 174 301
```

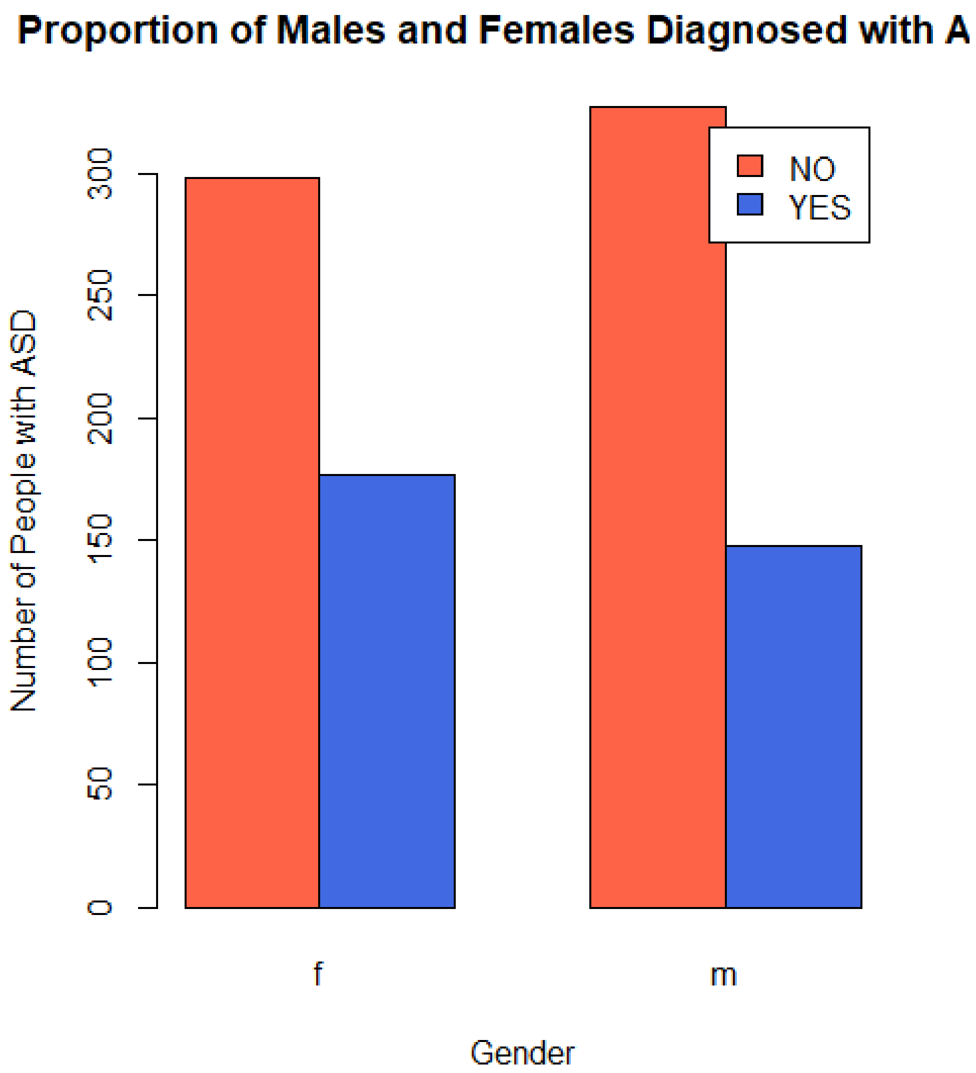
```
table(QMale$Q10)
```

```
##
##  no yes
## 199 276
```

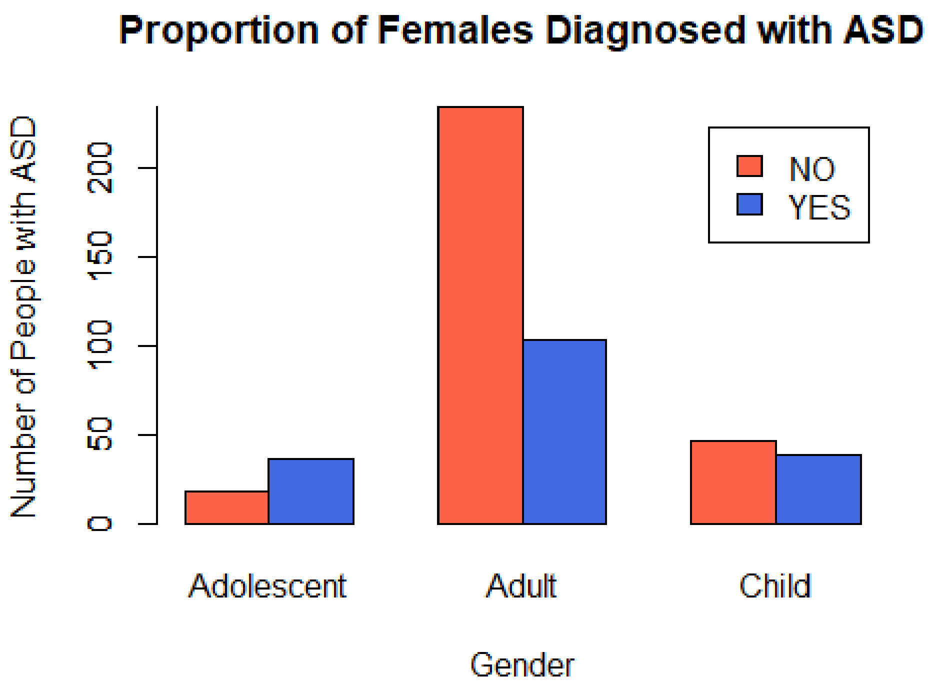
```
femalequestions <- data.frame(
  Questions = c(1, 1, 2, 2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10),
  Sample = c("Female", "Male", "Female", "Male", "Female", "Male", "Fem
ale", "Male", "Female", "Male", "Female", "Male", "Female", "Male", "Fe
male", "Male","Female", "Male", "Female", "Male"),
  No = c(122, 144, 242, 249, 227, 223, 213, 235, 199, 210, 266, 289, 27
2, 235, 198, 169, 275, 292, 174, 198),
  Yes = c(353, 329, 233, 224, 248, 250, 262, 238, 276, 263, 209, 184, 2
03, 238, 277, 304, 200, 181, 301, 275)
)

mfemalequestions <- melt(femalequestions, id.vars = 1:2)
ggplot(mfemalequestions, aes(x = Questions, y = value, fill = variabl
e), ordered=TRUE) +
  geom_bar(stat = "identity") +
  facet_grid(~Sample)
```



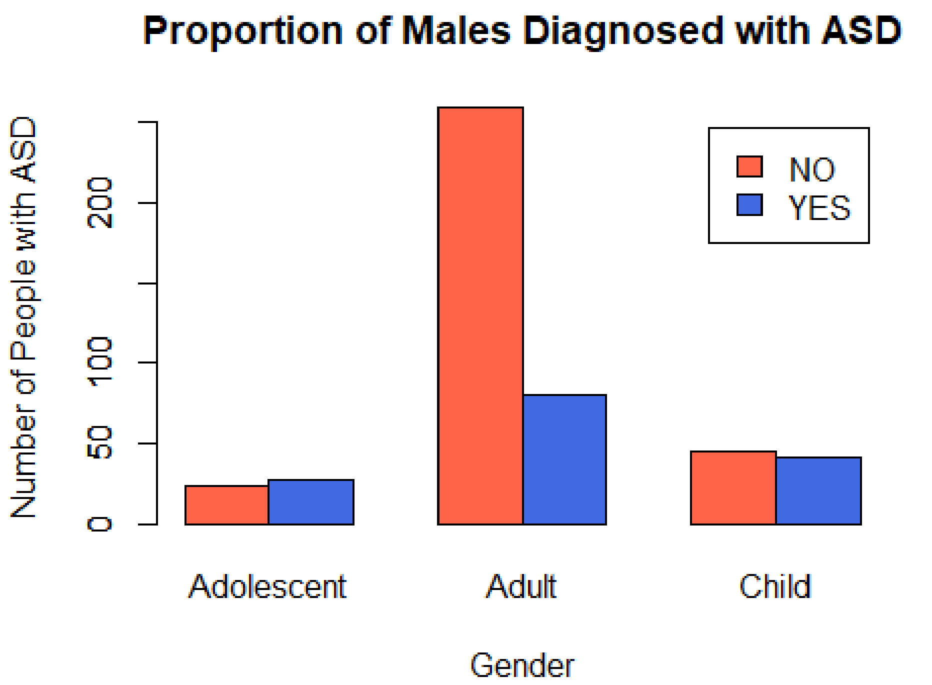

```
barplot(table(autism$ASD,autism$Gender), beside = TRUE, legend.text = TRUE, ylab = "Number of People with ASD", xlab = "Gender", col = c( "tomato", "royalblue"), main = "Proportion of Males and Females Diagnosed with ASD")
```




```
barplot(table(female$ASD,female$lifeStage), beside = TRUE, legend.text
= TRUE, ylab = "Number of People with ASD", xlab = "Gender", col = c( "
tomato", "royalblue"), main = "Proportion of Females Diagnosed with ASD
")
axis(2,at=seq(0,250,50))
```




```r
barplot(table(male$ASD,male$lifeStage), beside = TRUE, legend.text = TRUE, ylab = "Number of People with ASD", xlab = "Gender", col = c( "tomato", "royalblue"), main = "Proportion of Males Diagnosed with ASD")
axis(2,at=seq(0,250,50))
```

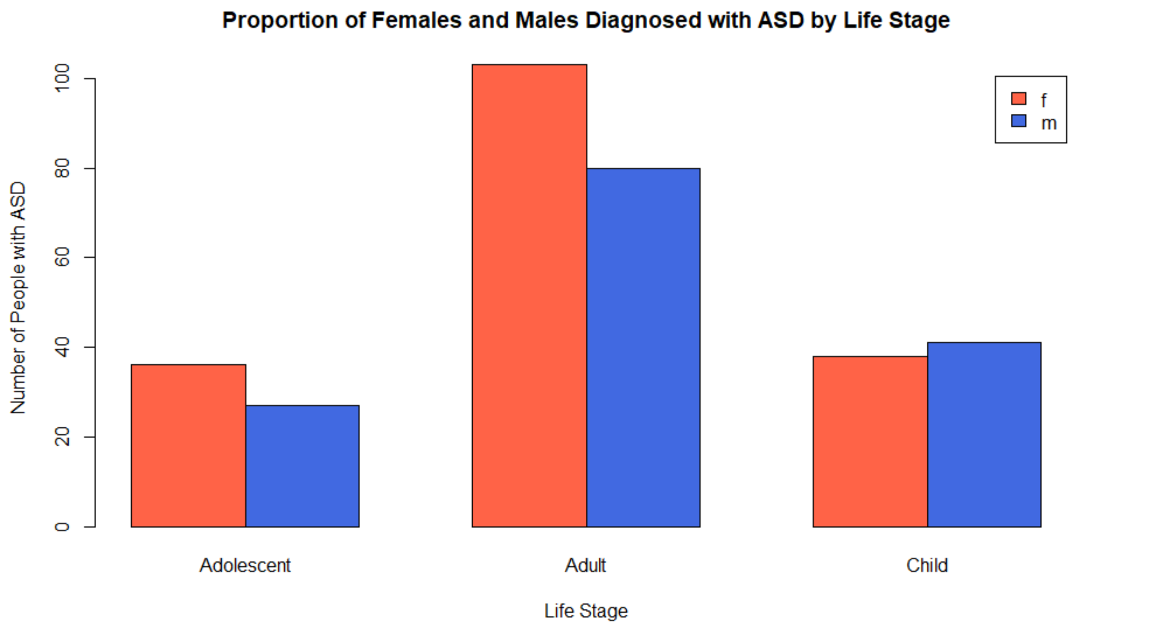


1st Hypothesis

```
maleswithASD <- subset(male, subset = ASD == "YES")
head(maleswithASD)

##     Q1  Q2  Q3  Q4  Q5  Q6  Q7  Q8  Q9 Q10 Age Gender              R
ace
## 20 yes yes yes  no yes yes yes  no yes yes  12      m    White-Europ
ean
## 22 yes yes yes yes  no yes yes yes yes yes  14      m    White-Europ
ean
## 24 yes  no yes yes yes yes yes yes yes yes  13      m 'Middle Easter
n '
## 25 yes yes yes yes yes yes yes yes yes yes  14      m          Hispa
nic
## 30  no yes yes yes yes yes yes yes yes yes  16      m            Oth
ers
## 32 yes yes yes yes yes yes  no  no  no yes  13      m    White-Europ
ean
##    Jaundice FamilyPDD        Residence SecondUse ScreeningScore
## 20       no        no 'United Kingdom'        no              8
## 22       no        no    'New Zealand'        no              9
## 24       no       yes  'United States'        no              9
## 25       no        no        Argentina        no             10
## 30       no        no 'United Kingdom'        no              9
## 32       no        no 'United Kingdom'        no              7
##          AgeRange Response ASD  lifeStage
## 20 '12-16 years'   Parent YES Adolescent
## 22 '12-16 years'   Parent YES Adolescent
## 24 '12-16 years'   Parent YES Adolescent
## 25 '12-16 years'     Self YES Adolescent
## 30 '12-16 years'   Parent YES Adolescent
## 32 '12-16 years'     Self YES Adolescent

femaleswithASD <- subset(female, subset = ASD == "YES")
head(femaleswithASD)

##     Q1  Q2  Q3  Q4  Q5  Q6  Q7  Q8  Q9 Q10 Age Gender              R
ace
## 4   no yes yes yes yes yes  no yes yes  no  14      f    White-Europ
ean
## 5  yes yes yes yes yes yes yes  no  no  no  16      f
  ?
## 8  yes yes  no yes yes  no yes yes  no yes  15      f 'Middle Easter
n '
## 13 yes  no  no yes yes yes  no yes yes yes  12      f            Oth
ers
## 14 yes yes yes yes yes yes  no yes yes yes  12      f            Oth
ers
## 15 yes  no yes yes yes yes  no yes yes yes  12      f            Oth
ers
##    Jaundice FamilyPDD        Residence SecondUse ScreeningScore
## 4        no        no 'United Kingdom'        no              7
## 5        no        no          Albania        no              7
## 8        no        no        Australia        no              7
```

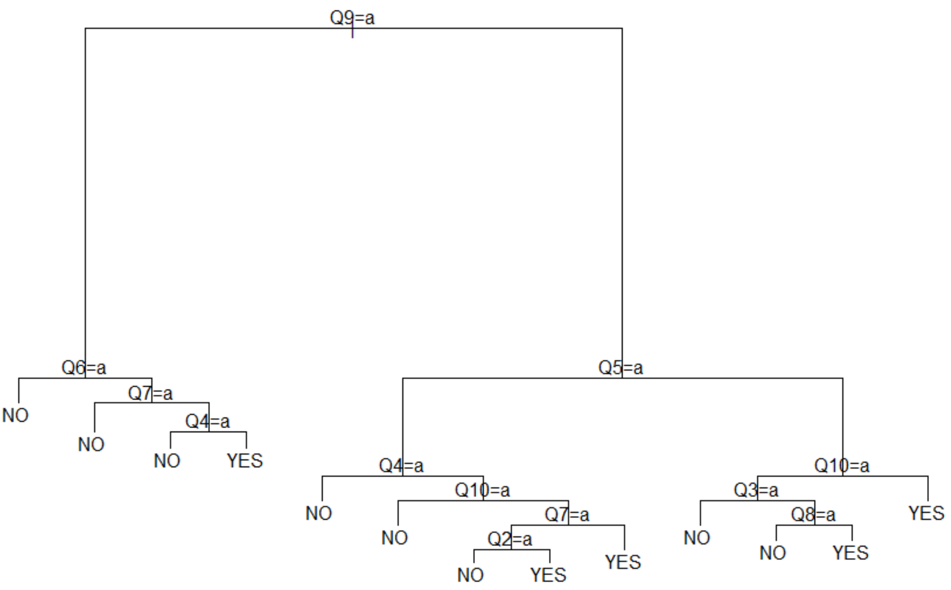
```
## 13        no         no 'United Kingdom'        no               7
## 14        no         no 'United Kingdom'        no               9
## 15        no         no 'United Kingdom'        no               8
##         AgeRange Response ASD  lifeStage
## 4  '12-16 years'     Self YES Adolescent
## 5  '12-16 years'        ? YES Adolescent
## 8  '12-16 years'   Parent YES Adolescent
## 13 '12-16 years'     Self YES Adolescent
## 14 '12-16 years'   Parent YES Adolescent
## 15 '12-16 years'     Self YES Adolescent
```

```
yestoautism <- rbind(femaleswithASD, maleswithASD)

barplot(table(yestoautism$Gender, yestoautism$lifeStage), beside = TRU
E, legend.text = TRUE, ylab = "Number of People with ASD", xlab = "Life
 Stage", col = c( "tomato", "royalblue"), main = "Proportion of Females
 and Males Diagnosed with ASD by Life Stage")
```

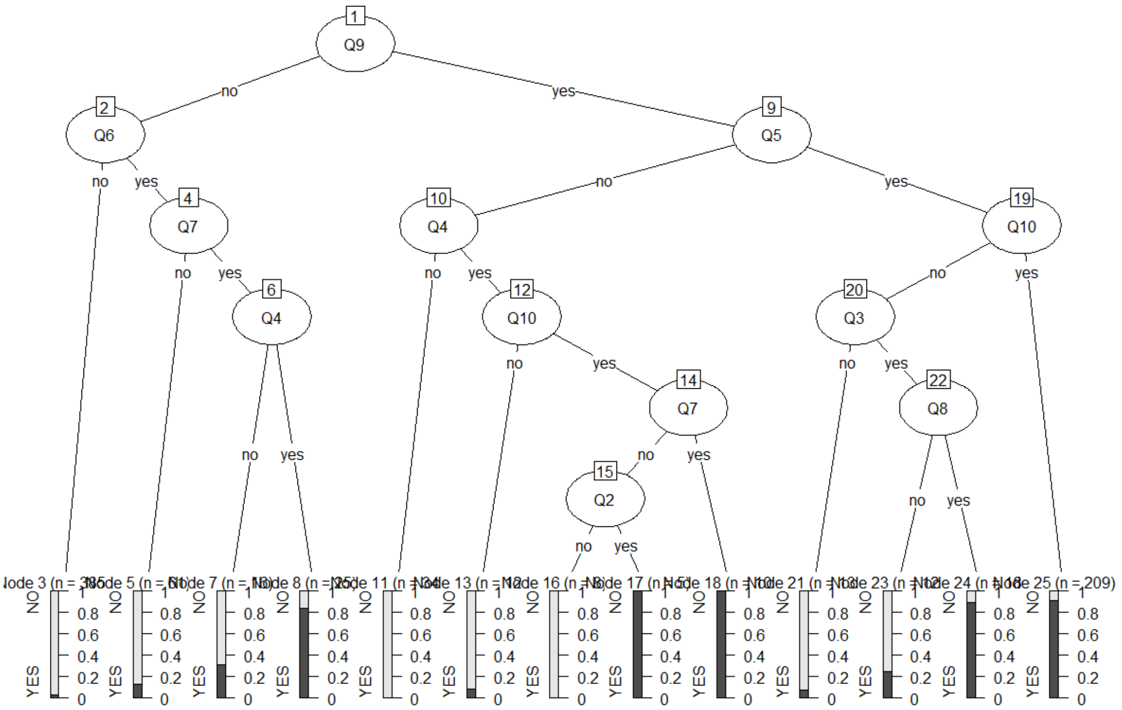


Random Forest

```
set.seed(54321)
Index <- sample(nrow(autism), floor(0.15 * nrow(autism)), replace = FAL
SE)
Trainautism <- autism[-Index, ]
Testautism <- autism[Index, ]

autism.rpart <- rpart(ASD~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9 +
 Q10, data=Trainautism, method = "class", minsplit = 2, minbucket = 1)

plot(autism.rpart)
text(autism.rpart)
```




```r
plot(as.party(autism.rpart))
```




```r
head(predict(autism.rpart, newdata = Testautism))
```

```
##             NO        YES
## 836  0.0861244 0.91387560
## 1098 0.1600000 0.84000000
## 370  0.9714286 0.02857143
```

```
## 563  0.0861244 0.91387560
## 448  0.9714286 0.02857143
## 694  0.9714286 0.02857143
```

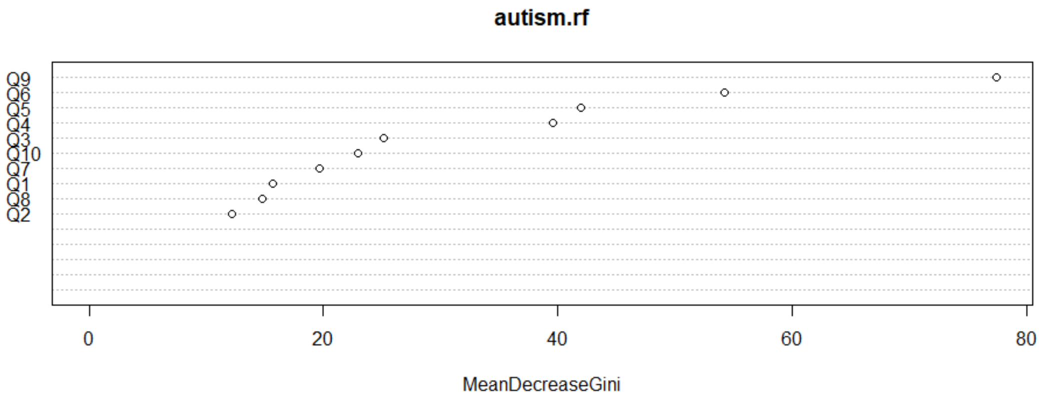
```
head(predict(autism.rpart, newdata = Testautism, type = "class"))
```

```
##  836 1098  370  563  448  694
##  YES  YES   NO  YES   NO   NO
## Levels: NO YES
```

```
autism.rpart.pred <- predict(autism.rpart, newdata = Testautism, type =
 "class")

table(autism.rpart.pred, Testautism$ASD, dnn = c("Predictions", "Actual
"))
```

```
##            Actual
## Predictions NO YES
##         NO  81  12
##         YES  8  41
```

```
autism.rf <- randomForest(ASD ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 +
 Q9 + Q10, data = Trainautism, importance=TRUE)

varImpPlot(autism.rf, type = 2, n.var = 15)
```




```
autism.rf.pred <- predict(autism.rf, newdata = Testautism)
table(autism.rf.pred, Testautism$ASD, dnn = c("Predictions", "Actual"))
```

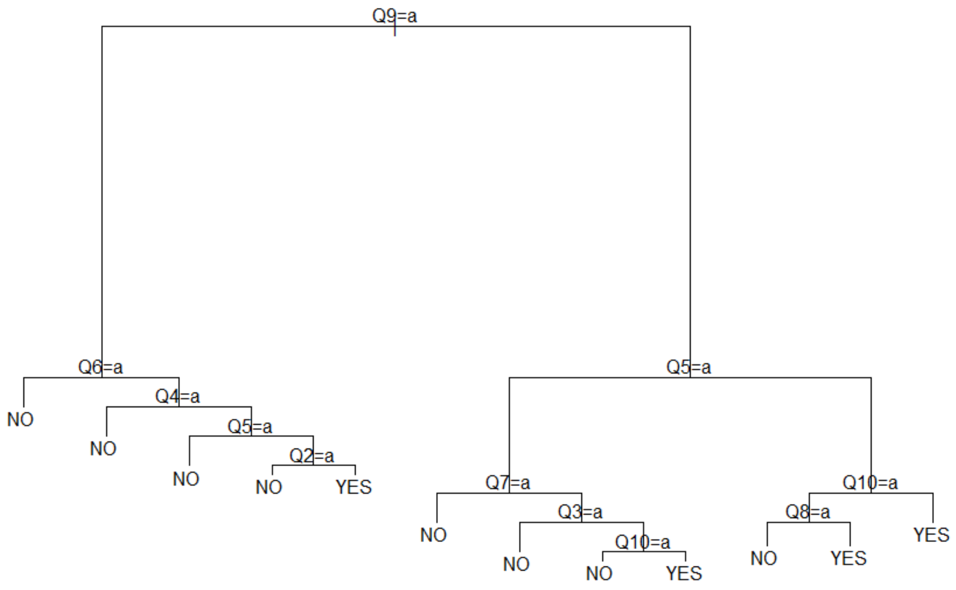
```
##            Actual
## Predictions NO YES
##         NO  88   3
##         YES  1  50
```

2nd Hypothesis

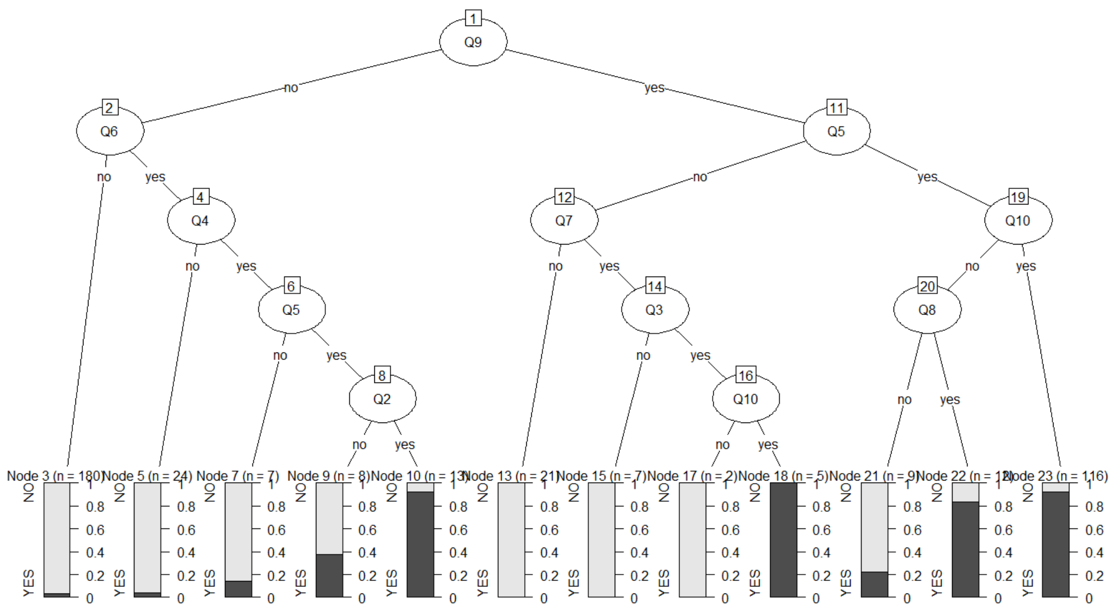
1) For female

```
set.seed(54321)
Index <- sample(nrow(female), floor(0.15 * nrow(female)), replace = FAL
SE)
```

```
Trainautism_f <- female[-Index, ]
Testautism_f <- female[Index, ]

autism_f.rpart <- rpart(ASD~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9
 + Q10, data=Trainautism_f, method = "class", minsplit = 2, minbucket =
 1)

plot(autism_f.rpart)
text(autism_f.rpart)
```



```
plot(as.party(autism_f.rpart))
```



```r
head(predict(autism_f.rpart, newdata = Testautism_f))
```

```
##             NO        YES
## 415 0.96666667 0.03333333
## 499 0.95833333 0.04166667
## 161 0.07758621 0.92241379
## 253 0.95833333 0.04166667
## 195 0.96666667 0.03333333
## 857 0.07758621 0.92241379
```

```r
head(predict(autism_f.rpart, newdata = Testautism_f, type = "class"))
```

```
## 415 499 161 253 195 857
##  NO  NO YES  NO  NO YES
## Levels: NO YES
```

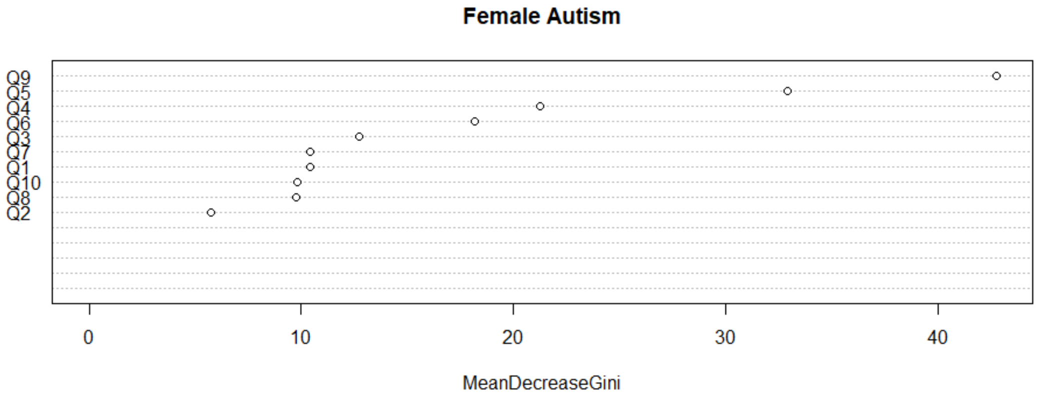
```r
autism_f.rpart.pred <- predict(autism_f.rpart, newdata = Testautism_f,
type = "class")
```

```r
table(autism_f.rpart.pred, Testautism_f$ASD, dnn = c("Predictions", "Ac
tual"))
```

```
##            Actual
## Predictions NO YES
##         NO  35   8
##         YES  6  22
```

```r
autism_f.rf <- randomForest(ASD ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8
 + Q9 + Q10, data = Trainautism_f)
```

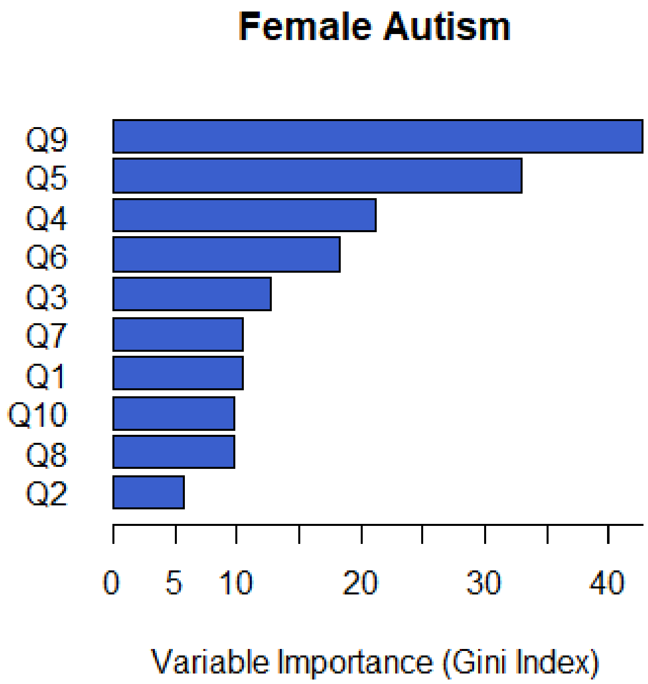
```r
varImpPlot(autism_f.rf, type = 2, n.var = 15, main = "Female Autism")
```




```r
autism_f.rf.pred <- predict(autism_f.rf, newdata = Testautism_f)
table(autism_f.rf.pred, Testautism_f$ASD, dnn = c("Predictions", "Actua
l"))
```

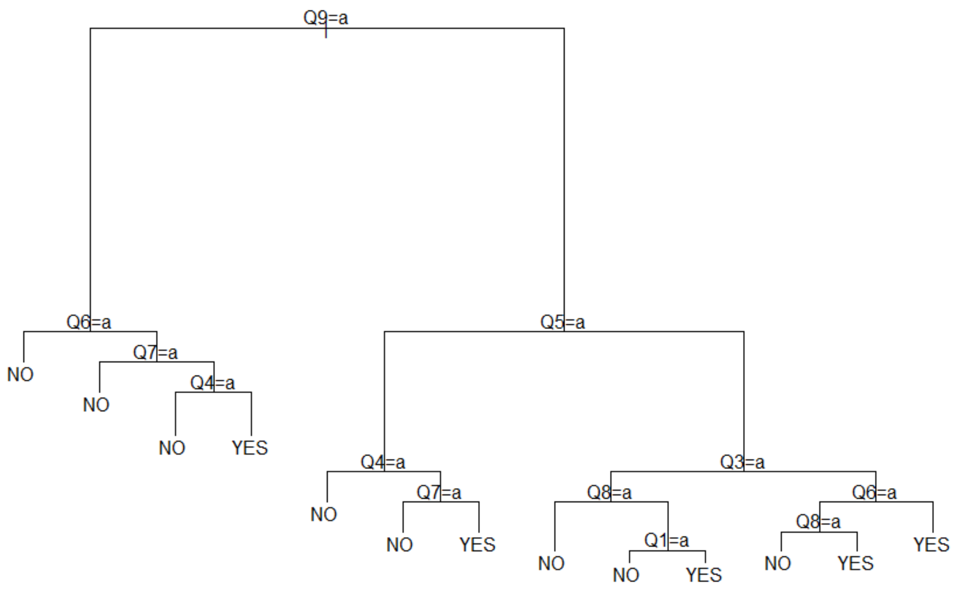
```
##            Actual
## Predictions NO YES
##         NO  40   4
##         YES  1  26
```

```r
autism_f.rf.imp <- autism_f.rf$importance
autism_f.rf.imp10 <- sort(autism_f.rf.imp[, ], decreasing = TRUE)[1:10]

par(oma = c(0,5,0,0))
par(las=2)
barplot(rev(autism_f.rf.imp10), horiz = TRUE, col = "royalblue3", xlab=
"Variable Importance (Gini Index)",xaxt="n", main = "Female Autism")
axis(side=1, at = c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50), labels =
c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50), las = 1)
```

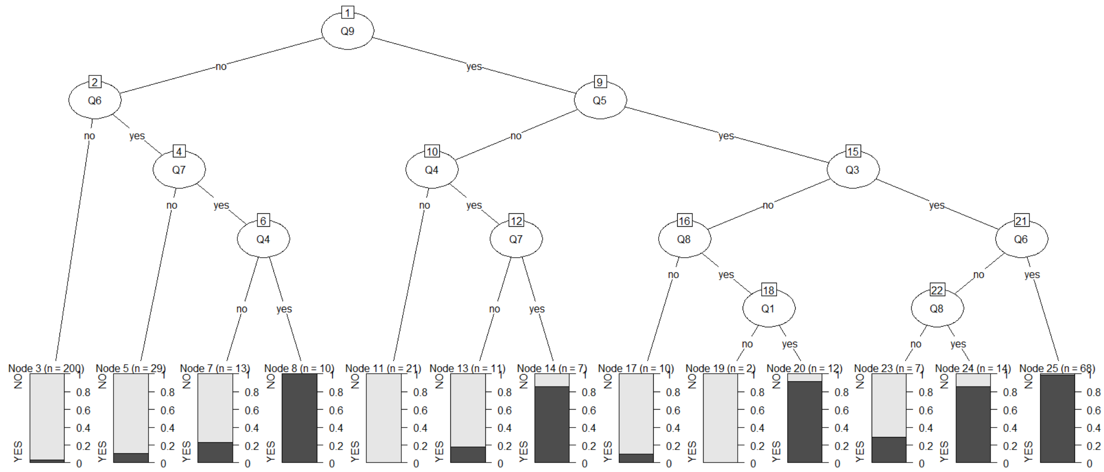


2) For Males

```r
set.seed(54321)
Index <- sample(nrow(male), floor(0.15 * nrow(male)), replace = FALSE)
Trainautism_m <- male[-Index, ]
Testautism_m <- male[Index, ]

autism_m.rpart <- rpart(ASD~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9
 + Q10, data=Trainautism_m, method = "class", minsplit = 2, minbucket =
 1)

plot(autism_m.rpart)
text(autism_m.rpart)
```




```
#fancyRpartPlot(autism.rpart)

plot(as.party(autism_m.rpart))
```




```
head(predict(autism_m.rpart, newdata = Testautism_m))

##              NO       YES
## 446 0.96500000 0.0350000
## 516 0.96500000 0.0350000
## 239 1.00000000 0.0000000
## 323 0.96500000 0.0350000
## 265 0.96500000 0.0350000
## 836 0.01470588 0.9852941

head(predict(autism_m.rpart, newdata = Testautism_m, type = "class"))
```

```
## 446 516 239 323 265 836
##  NO  NO  NO  NO  NO YES
## Levels: NO YES
```

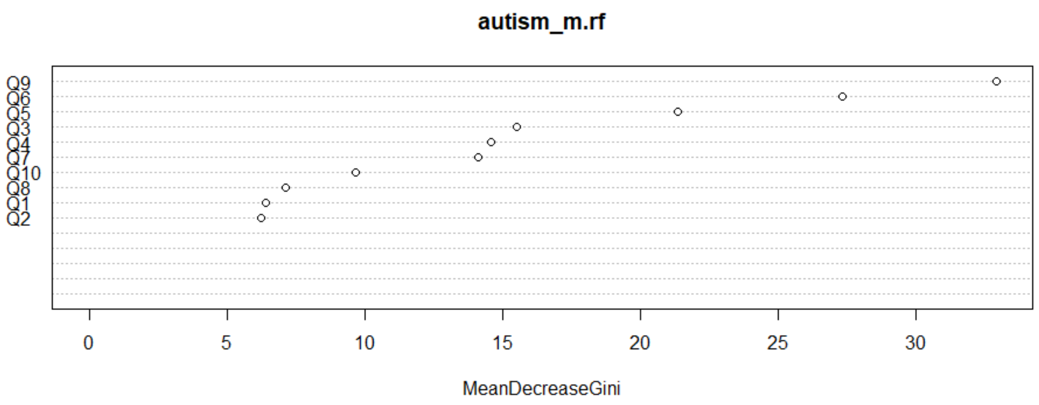
```
autism_m.rpart.pred <- predict(autism_m.rpart, newdata = Testautism_m,
type = "class")
```

```
table(autism_m.rpart.pred, Testautism_m$ASD, dnn = c("Predictions", "Ac
tual"))
```

```
##            Actual
## Predictions NO YES
##         NO  45   5
##         YES  2  19
```

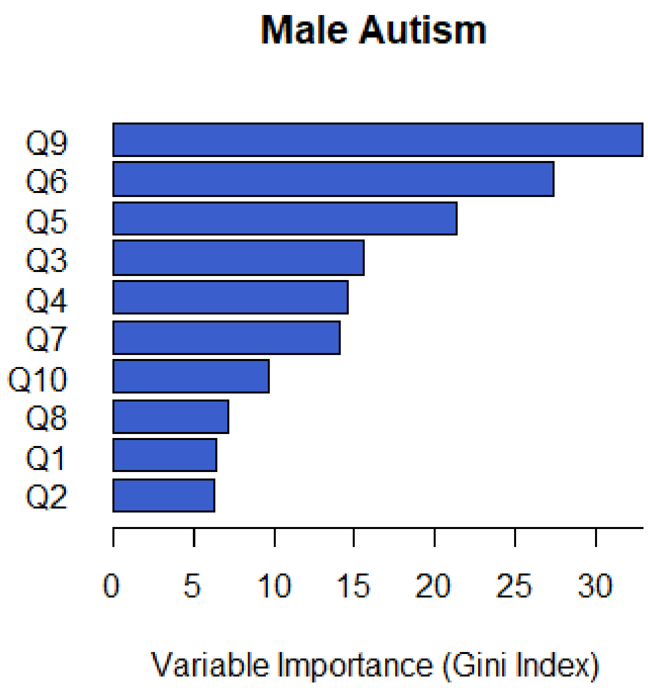
```
autism_m.rf <- randomForest(ASD ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8
 + Q9 + Q10, data = Trainautism_m)
```

```
varImpPlot(autism_m.rf, type = 2, n.var = 15)
```




```
autism_m.rf.imp <- autism_m.rf$importance
autism_m.rf.imp10 <- sort(autism_m.rf.imp[, ], decreasing = TRUE)[1:10]
```

```
par(oma = c(0,5,0,0))
par(las=2)
barplot(rev(autism_m.rf.imp10), horiz = TRUE, col = "royalblue3", xlab=
"Variable Importance (Gini Index)",xaxt="n", main = "Male Autism")
axis(side=1, at = c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50), labels =
c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50), las = 1)
```




```
autism_m.rf.pred <- predict(autism_m.rf, newdata = Testautism_m)
table(autism_m.rf.pred, Testautism_m$ASD, dnn = c("Predictions", "Actua
l"))

##            Actual
## Predictions NO YES
##         NO  46   1
##         YES  1  23
```

Using logistic regression

```
glmall <- glm(ASD ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9 + Q10, d
ata = Trainautism, family = binomial(logit))

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(glmall)

##
## Call:
## glm(formula = ASD ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9 +
##     Q10, family = binomial(logit), data = Trainautism)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.907e-05  -2.110e-08  -2.110e-08   2.110e-08   2.553e-05
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  -290.27   44762.22  -0.006    0.995
## Q1yes           44.96   10767.08   0.004    0.997
## Q2yes           44.41    9518.91   0.005    0.996
## Q3yes           44.23    9599.50   0.005    0.996
## Q4yes           44.79    9696.04   0.005    0.996
## Q5yes           44.95   10124.67   0.004    0.996
## Q6yes           44.74    9416.34   0.005    0.996
## Q7yes           44.53    9411.97   0.005    0.996
## Q8yes           44.61    9802.71   0.005    0.996
## Q9yes           44.65    9129.41   0.005    0.996
## Q10yes          44.60    9888.03   0.005    0.996
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.0323e+03  on 807  degrees of freedom
## Residual deviance: 6.3093e-08  on 797  degrees of freedom
## AIC: 22
##
## Number of Fisher Scoring iterations: 25
```

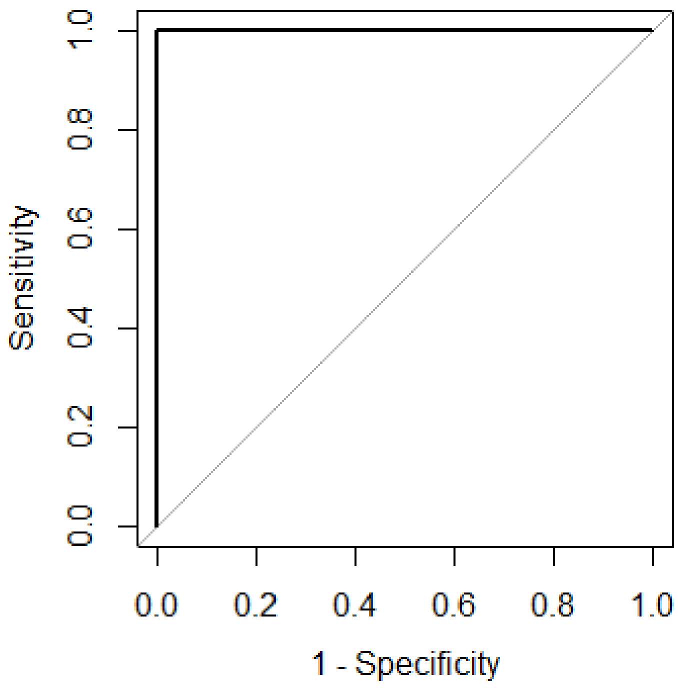
```
allpred <- predict(glmall, newdata = Testautism, type = "response")
head(allpred)
```

```
##          836         1098          370          563          448
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 2.220446e-16
##          694
## 2.220446e-16
```

```
allpred.class <- rep("NO", length(allpred))
allpred.class[allpred > 0.5] <- "YES"

table(allpred.class, Testautism$ASD)
```

```
##
## allpred.class NO YES
##           NO  89   0
##           YES  0  53
```

```
par(pty = "s")
plot(roc(Testautism$ASD, allpred), legacy.axes = TRUE)
```




```
glmall1 <- glm(ASD ~ ScreeningScore, data = Trainautism, family = binom
ial(logit))

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(glmall1)

##
## Call:
## glm(formula = ASD ~ ScreeningScore, family = binomial(logit),
##     data = Trainautism)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.002e-05  -2.110e-08  -2.110e-08   2.110e-08   1.952e-05
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -290.60   44295.21  -0.007    0.995
## ScreeningScore     44.71    6827.33   0.007    0.995
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.0323e+03  on 807  degrees of freedom
## Residual deviance: 6.3386e-08  on 806  degrees of freedom
## AIC: 4
##
## Number of Fisher Scoring iterations: 25
```

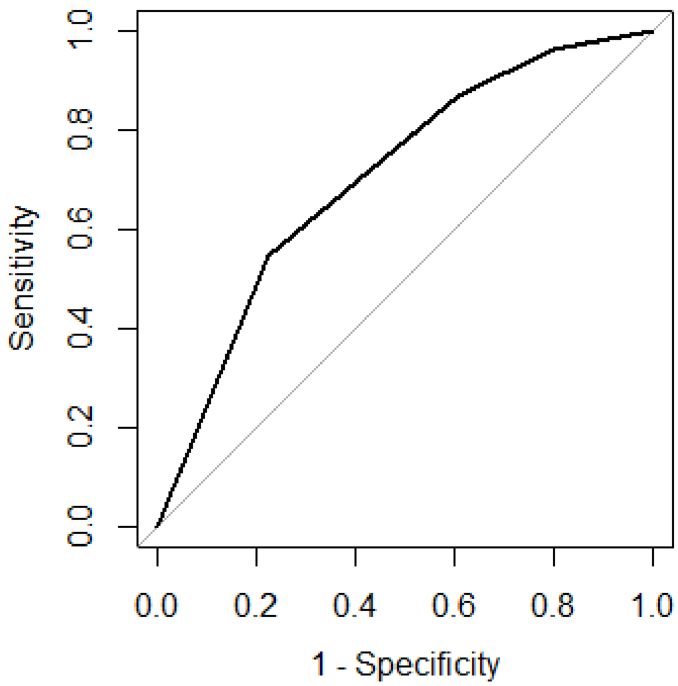
```r
glm2 <- glm(ASD ~ Q1 + Q2, data = Trainautism, family = binomial(logi
t))
summary(glm2)
```

```
##
## Call:
## glm(formula = ASD ~ Q1 + Q2, family = binomial(logit), data = Traina
utism)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3191  -0.7851  -0.6674   1.0420   2.3654
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.7347     0.2370 -11.541  < 2e-16 ***
## Q1yes         1.7156     0.2263   7.581 3.43e-14 ***
## Q2yes         1.3461     0.1655   8.136 4.09e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1032.26  on 807  degrees of freedom
## Residual deviance:  890.92  on 805  degrees of freedom
## AIC: 896.92
##
## Number of Fisher Scoring iterations: 4
```

```r
Q2pred <- predict(glm2, newdata = Testautism, type = "response")
Q2pred.class <- rep("NO", length(Q2pred))
Q2pred.class[Q2pred > 0.55] <- "YES"
table(Q2pred.class, Testautism$ASD)
```

```
##
## Q2pred.class NO YES
##          NO  69  24
##          YES 20  29
```

```r
par(pty = "s")
plot(roc(Testautism$ASD, Q2pred), legacy.axes = TRUE)
```


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity

6. References

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (DSM-5Â®). American Psychiatric Pub.

Johnson, C. P., & Myers, S. M. (2007). Identification and evaluation of children with autism spectrum disorders. Pediatrics, 120(5), 1183-1215.

Kopp, S., & Gillberg, C. (2011). The Autism Spectrum Screening Questionnaire (ASSQ)-Revised Extended Version (ASSQ-REV): an instrument for better capturing the autism phenotype in girls? A preliminary study involving 191 clinical cases and community controls. Research in developmental disabilities, 32(6), 2875-2888.

Lai, M. C., Lombardo, M. V., Auyeung, B., Chakrabarti, B., & Baron-Cohen, S. (2015). Sex/gender differences and autism: setting the scene for future research. Journal of the American Academy of Child & Adolescent Psychiatry, 54(1), 11-24.

Mandy, W., Chilvers, R., Chowdhury, U., Salter, G., Seigal, A., & Skuse, D. (2012). Sex differences in autism spectrum disorder: evidence from a large sample of children and adolescents. Journal of autism and developmental disorders, 42(7), 1304-1313.

Newschaffer, C. J., Croen, L. A., Daniels, J., Giarelli, E., Grether, J. K., Levy, S. E., … & Reynolds, A. M. (2007). The epidemiology of autism spectrum disorders. Annu. Rev. Public Health, 28, 235-258.

Thabtah, F. (2017). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, 1-6.

Thabtah, F. (2018). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics for Health and Social Care, 1-20.

Worley, J. A., & Matson, J. L. (2012). Comparing symptoms of autism spectrum disorders using the current DSM-IV-TR diagnostic criteria and the proposed DSM-V diagnostic criteria. Research in Autism Spectrum Disorders, 6(2), 965-970.