Exploratory Data Analysis for ASD-Toddler dataset

Exploratory Data Analysis (EDA) for univariate variables

Dataset

Autism Spectrum Disorder for Toddlers

Description

Use for Autism Screening contributed by Dr. Fadi Fayez.

UCI: https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children+
+

(this is not the exact dataset rather related with same variable descriptors but differentiates with the dimensionality)

Use funModeling package developed by Pablo Casas

Problem:

Use Supervised Learning Techniques such as rule-based classifiers to predict the outcome of a categorical variable "Class" on whether a toddler is positively labelled with Autism traits or not.

Start of EDA

```
# Install Required Packages
#install.packages("tidyverse", repos = "http://cran.us.r-project.org")
#install.packages("funModeling", repos = "http://cran.us.r-project.org") # EDA tool
#install.packages("Hmisc", repos = "http://cran.us.r-project.org") # gives an overview of all the
#variables in tabular format
#install.packages("FREQ", repos = "http://cran.us.r-project.org")

library(knitr)
library(funModeling)
```

```
## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
## funModeling v.1.6.8 :)
## Examples and tutorials at livebook.datascienceheroes.com
```

```r
library(Hmisc)
library(FREQ)
```

```
##
## Attaching package: 'FREQ'
## The following object is masked from 'package:funModeling':
##
##     freq
```

```r
library(tibble)

# load asd toddler dataset
asd <- read.csv("/Users/rmph/Desktop/Projects - current/Project ADA.A/dataset/toddler.csv")

# Start Profiling Categorical Variables
# Report missing values, descriptive statistics
describe(asd) # numerical and categorical profiling (quantitative)
```

```
## asd
##
##  18  Variables      1054  Observations
## ---------------------------------------------------------------------------
## A1
##        n  missing distinct     Info      Sum     Mean      Gmd
##     1054        0        2    0.738      594   0.5636   0.4924
##
## ---------------------------------------------------------------------------
## A2
##        n  missing distinct     Info      Sum     Mean      Gmd
##     1054        0        2    0.742      473   0.4488   0.4952
##
## ---------------------------------------------------------------------------
## A3
##        n  missing distinct     Info      Sum     Mean      Gmd
##     1054        0        2    0.721      423   0.4013    0.481
##
## ---------------------------------------------------------------------------
## A4
##        n  missing distinct     Info      Sum     Mean      Gmd
##     1054        0        2     0.75      540   0.5123   0.5002
##
## ---------------------------------------------------------------------------
## A5
##        n  missing distinct     Info      Sum     Mean      Gmd
##     1054        0        2    0.748      553   0.5247   0.4993
##
## ---------------------------------------------------------------------------
## A6
```

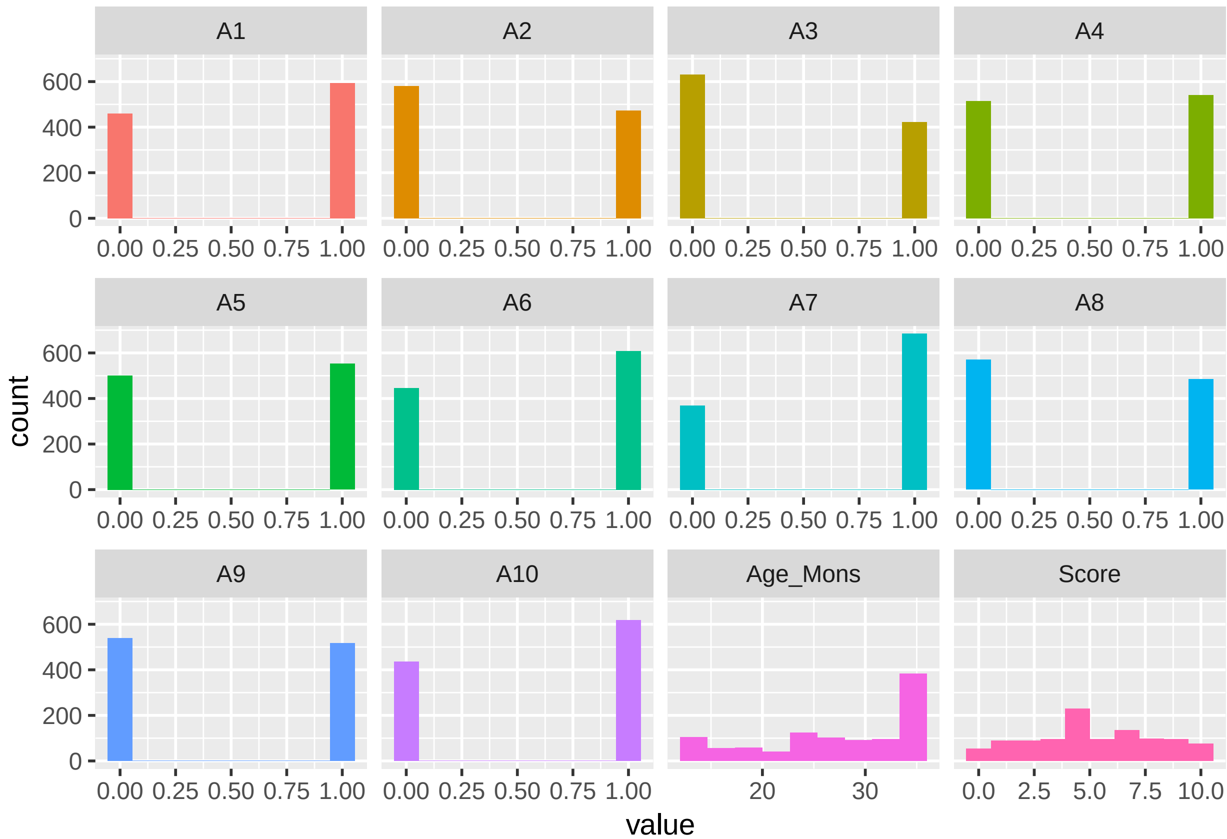
```
##        n  missing distinct     Info      Sum     Mean      Gmd 
##     1054        0        2    0.732      608   0.5769   0.4887 
## 
## --------------------------------------------------------------------------------
## A7 
##        n  missing distinct     Info      Sum     Mean      Gmd 
##     1054        0        2    0.683      685   0.6499   0.4555 
## 
## --------------------------------------------------------------------------------
## A8 
##        n  missing distinct     Info      Sum     Mean      Gmd 
##     1054        0        2    0.745      484   0.4592   0.4971 
## 
## --------------------------------------------------------------------------------
## A9 
##        n  missing distinct     Info      Sum     Mean      Gmd 
##     1054        0        2     0.75      516   0.4896   0.5003 
## 
## --------------------------------------------------------------------------------
## A10 
##        n  missing distinct     Info      Sum     Mean      Gmd 
##     1054        0        2    0.728      618   0.5863   0.4856 
## 
## --------------------------------------------------------------------------------
## Age_Mons 
##        n  missing distinct     Info     Mean      Gmd      .05      .10 
##     1054        0       25    0.971    27.87    8.859       12       15 
##      .25      .50      .75      .90      .95 
##       23       30       36       36       36 
## 
## lowest : 12 13 14 15 16, highest: 32 33 34 35 36
## --------------------------------------------------------------------------------
## Score 
##        n  missing distinct     Info     Mean      Gmd      .05      .10 
##     1054        0       11    0.991    5.213    3.338        0        1 
##      .25      .50      .75      .90      .95 
##        3        5        8        9       10 
## 
## Value          0     1     2     3     4     5     6     7     8     9
## Frequency     54    88    88    96   110   120    96   135    97    95
## Proportion 0.051 0.083 0.083 0.091 0.104 0.114 0.091 0.128 0.092 0.090
## 
## Value         10
## Frequency     75
## Proportion 0.071
## --------------------------------------------------------------------------------
## Sex 
##        n  missing distinct 
##     1054        0        2 
## 
## Value          f     m
## Frequency    319   735
## Proportion 0.303 0.697
## --------------------------------------------------------------------------------
```

```
## Ethnicity
##        n  missing distinct
##     1054        0       11
##
## asian (299, 0.284), black (53, 0.050), Hispanic (40, 0.038), Latino (26,
## 0.025), middle eastern (188, 0.178), mixed (8, 0.008), Native Indian (3,
## 0.003), Others (35, 0.033), Pacifica (8, 0.008), south asian (60, 0.057),
## White European (334, 0.317)
## ---------------------------------------------------------------------------
## Jauundice
##        n  missing distinct
##     1054        0        2
##
## Value         no   yes
## Frequency    766   288
## Proportion 0.727 0.273
## ---------------------------------------------------------------------------
## Family_ASD
##        n  missing distinct
##     1054        0        2
##
## Value         no   yes
## Frequency    884   170
## Proportion 0.839 0.161
## ---------------------------------------------------------------------------
## Who.completed.the.test
##        n  missing distinct
##     1054        0        5
##
## family member (1018, 0.966), Health care professional (5, 0.005), Health
## Care Professional (24, 0.023), Others (3, 0.003), Self (4, 0.004)
## ---------------------------------------------------------------------------
## Class
##        n  missing distinct
##     1054        0        2
##
## Value         No   Yes
## Frequency    326   728
## Proportion 0.309 0.691
## ---------------------------------------------------------------------------
```

Visualisation for Toddler-Dataset

```r
#freq(asd) # categorical variable profiling (quantitative and plot), path_out = "." (export plots)

plot_num(asd) # report distribution of numeric variables
```




```r
# Linear correlation: For all numerical variables only
# Pearson coefficient  (standard correlation measure)
# Score variable is the most important numerical variable, higher the value the higher the possibility
# Age_mons the lower the value and is less significant to the target class
correlation_table(asd, "Class")
```

```
##   Variable Class
## 1    Class  1.00
## 2    Score  0.81
## 3 Age_Mons  0.07
```

```r
# Calculates correlation of variables based on information theory metrics
# between the target class and the input variables
var_rank_info(asd, "Class")
```

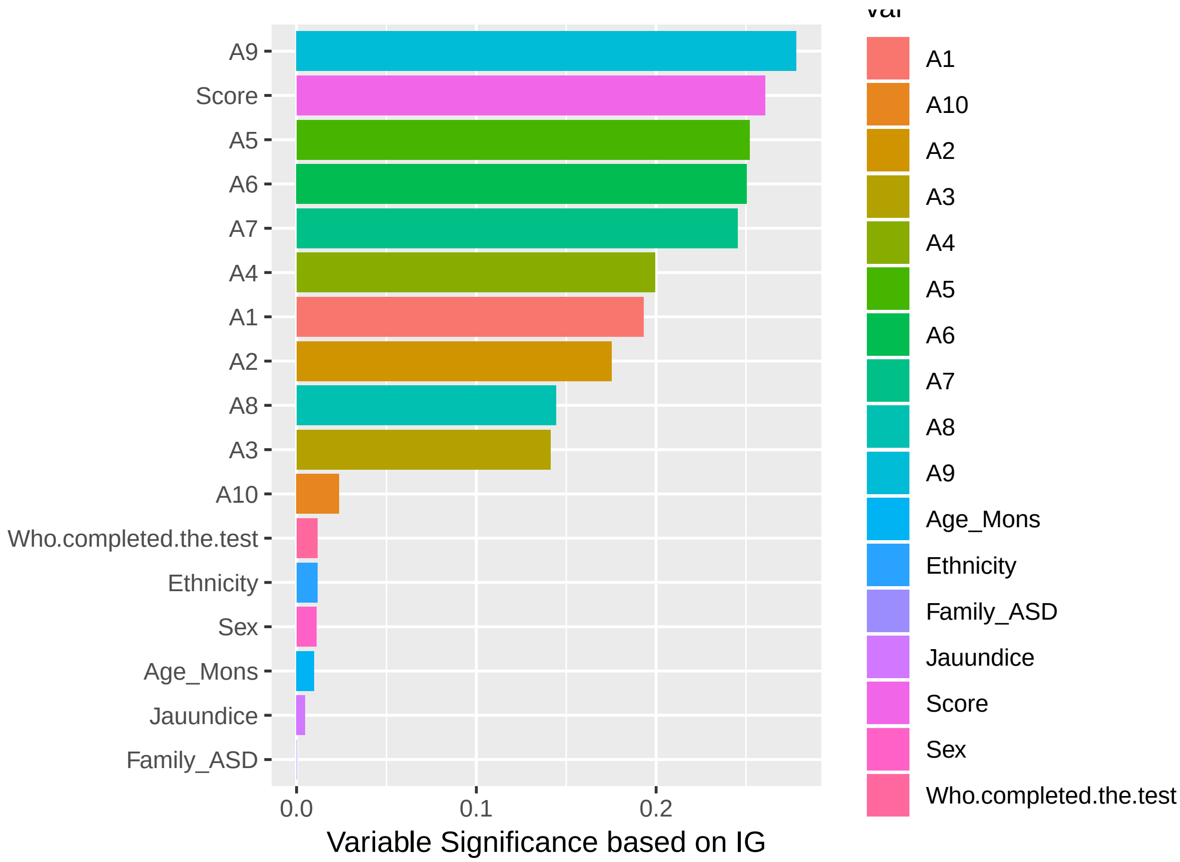
```
##                         var    en    mi          ig          gr
## 1                        A9 1.614 0.278 0.277909239 0.277996613
## 2                     Score 3.425 0.892 0.892361356 0.260515585
## 3                        A5 1.639 0.252 0.251609995 0.252052724
## 4                        A6 1.629 0.246 0.246159748 0.250444521
## 5                        A7 1.597 0.229 0.229182242 0.245337007
## 6                        A4 1.693 0.199 0.199315809 0.199403345
## 7                        A1 1.690 0.191 0.190788892 0.193045776
## 8                        A2 1.711 0.174 0.173790565 0.175119204
## 9                        A8 1.744 0.144 0.143630174 0.144324050
## 10                       A3 1.727 0.137 0.137305737 0.141301418
## 11                      A10 1.848 0.023 0.023145729 0.023657112
## 12 Who.completed.the.test 1.153 0.003 0.003077339 0.011661620
```

```
## 13              Ethnicity 3.405 0.030 0.029537696 0.011620373
## 14                    Sex 1.767 0.010 0.009769513 0.011045063
## 15               Age_Mons 4.810 0.039 0.038566543 0.009749532
## 16              Jauundice 1.734 0.004 0.004052141 0.004789299
## 17             Family_ASD 1.530 0.000 0.000130644 0.000204968
```

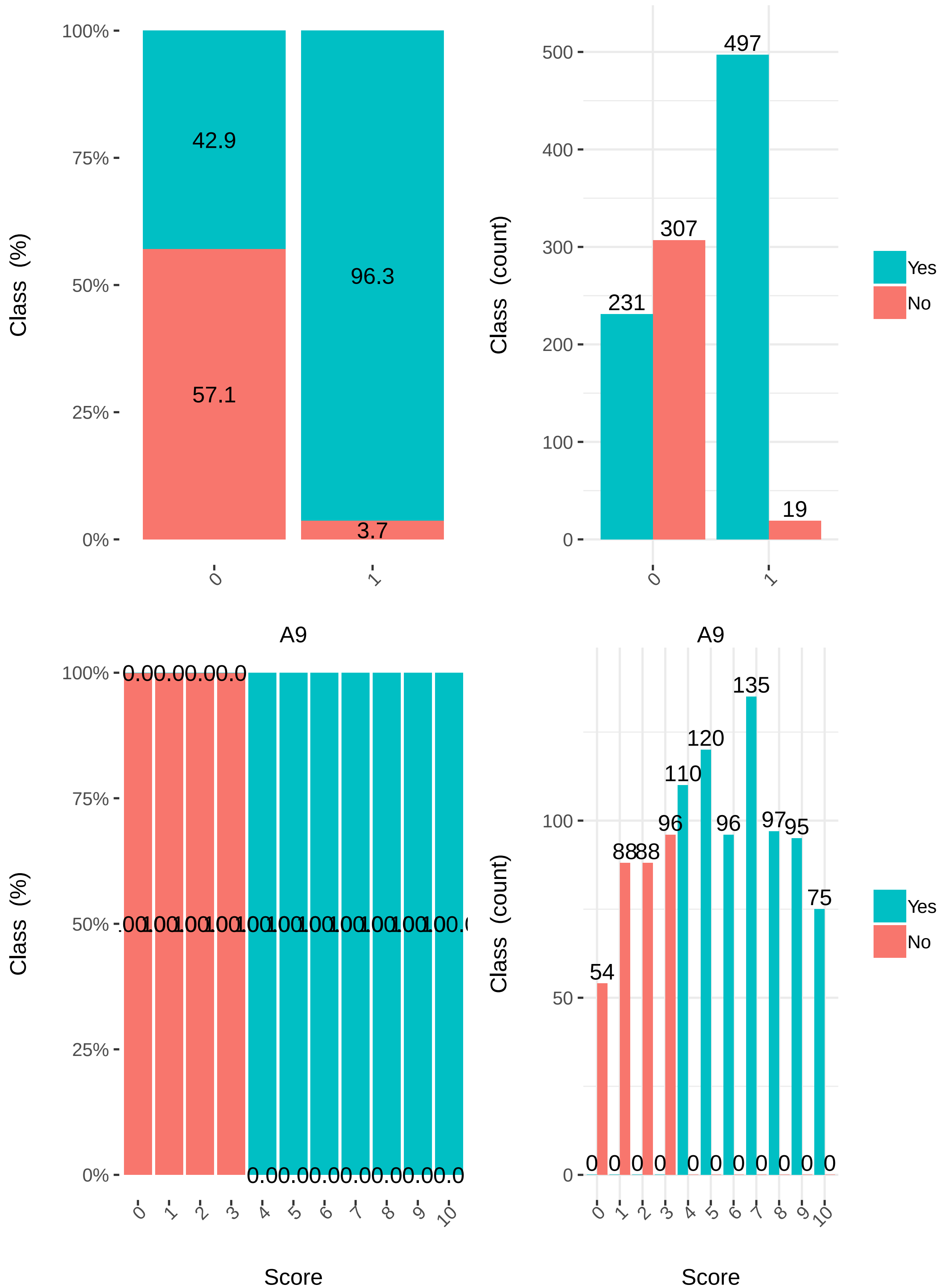
```r
sigvar <- var_rank_info(asd, target ="Class")
sigvar
```

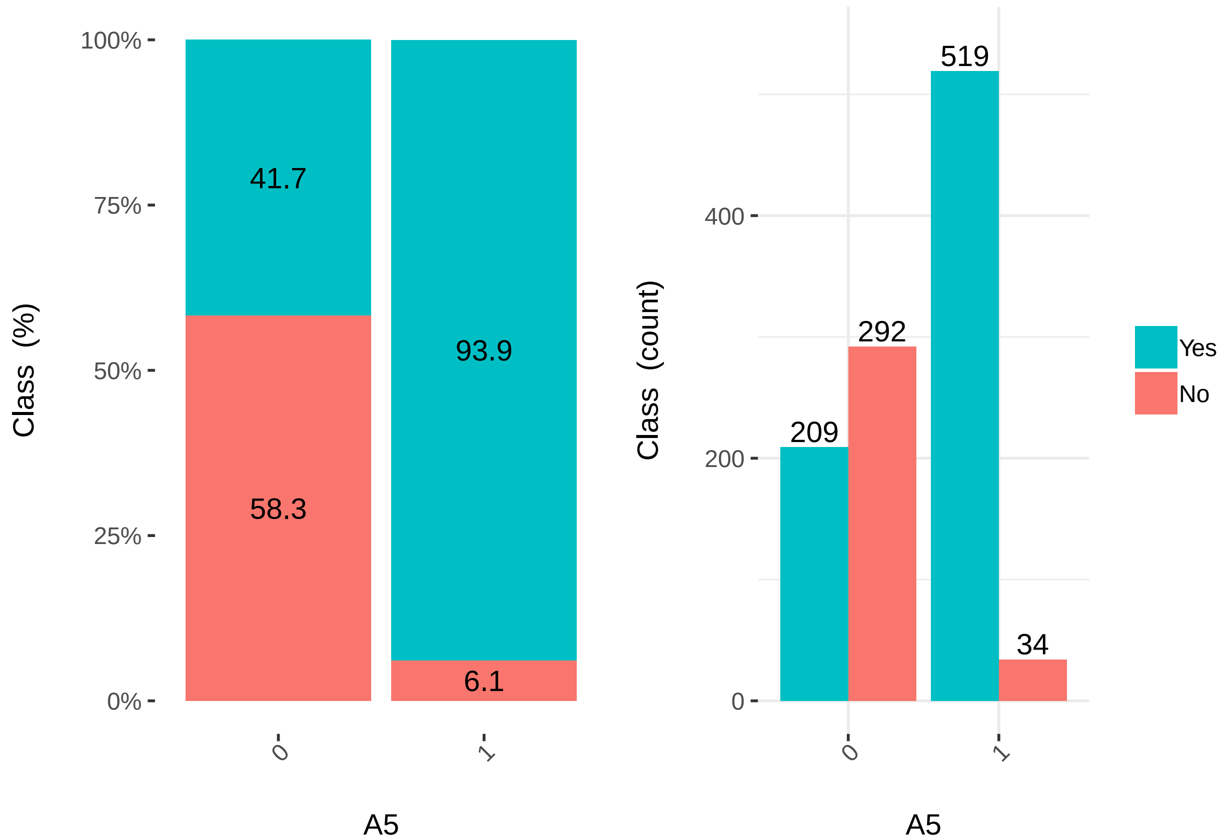
```
##                       var    en    mi          ig          gr
## 1                      A9 1.614 0.278 0.277909239 0.277996613
## 2                   Score 3.425 0.892 0.892361356 0.260515585
## 3                      A5 1.639 0.252 0.251609995 0.252052724
## 4                      A6 1.629 0.246 0.246159748 0.250444521
## 5                      A7 1.597 0.229 0.229182242 0.245337007
## 6                      A4 1.693 0.199 0.199315809 0.199403345
## 7                      A1 1.690 0.191 0.190788892 0.193045776
## 8                      A2 1.711 0.174 0.173790565 0.175119204
## 9                      A8 1.744 0.144 0.143630174 0.144324050
## 10                     A3 1.727 0.137 0.137305737 0.141301418
## 11                    A10 1.848 0.023 0.023145729 0.023657112
## 12 Who.completed.the.test 1.153 0.003 0.003077339 0.011661620
## 13              Ethnicity 3.405 0.030 0.029537696 0.011620373
## 14                    Sex 1.767 0.010 0.009769513 0.011045063
## 15               Age_Mons 4.810 0.039 0.038566543 0.009749532
## 16              Jauundice 1.734 0.004 0.004052141 0.004789299
## 17             Family_ASD 1.530 0.000 0.000130644 0.000204968
```

```r
#plotting variable significance
# the highest gr (gain ratio) is for variable A9 which maps to the QChat question and most relevant to
# The rest of the categorical variables are the ranked lowest by IG
r <- ggplot(data=sigvar, aes(x=reorder(var,gr), y=gr, fill=var))
r + geom_bar(stat = "identity") +
  coord_flip() +
  theme_get() +
  xlab("") +
  ylab("Variable Significance based on IG")
```

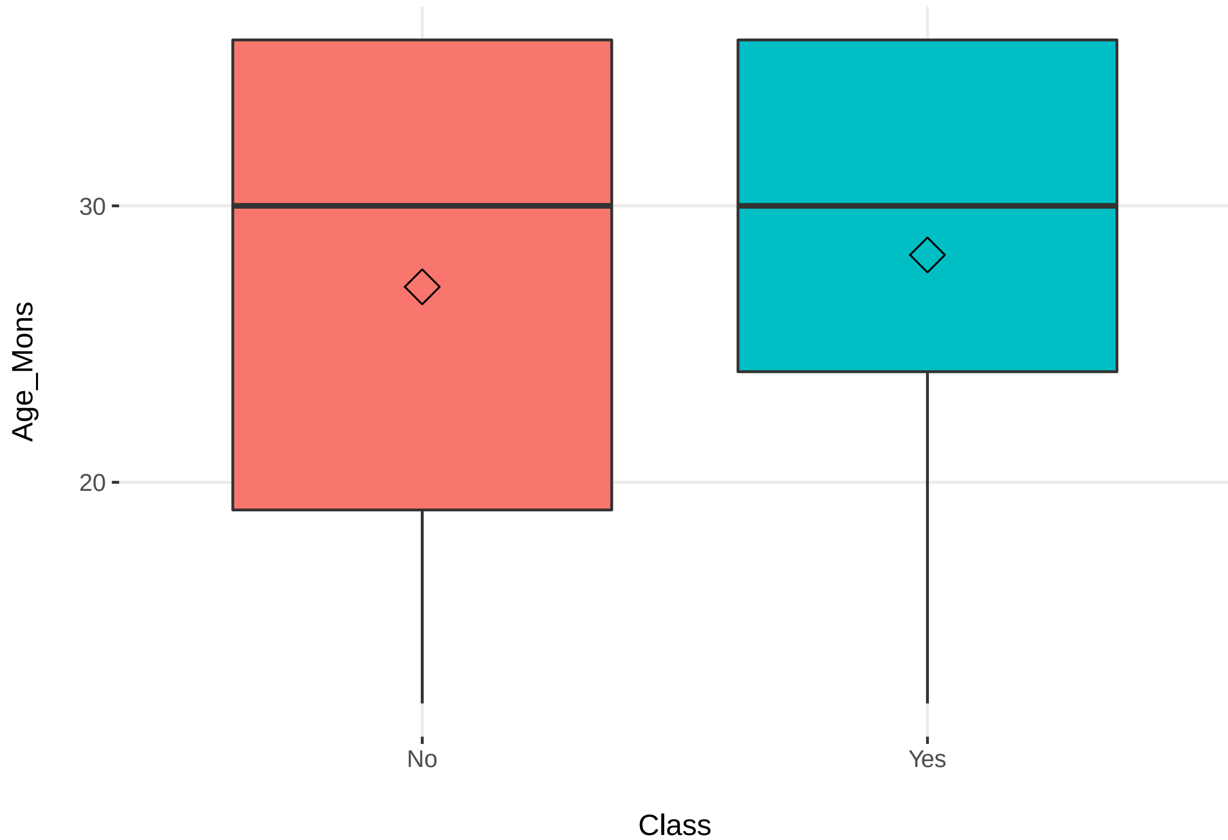


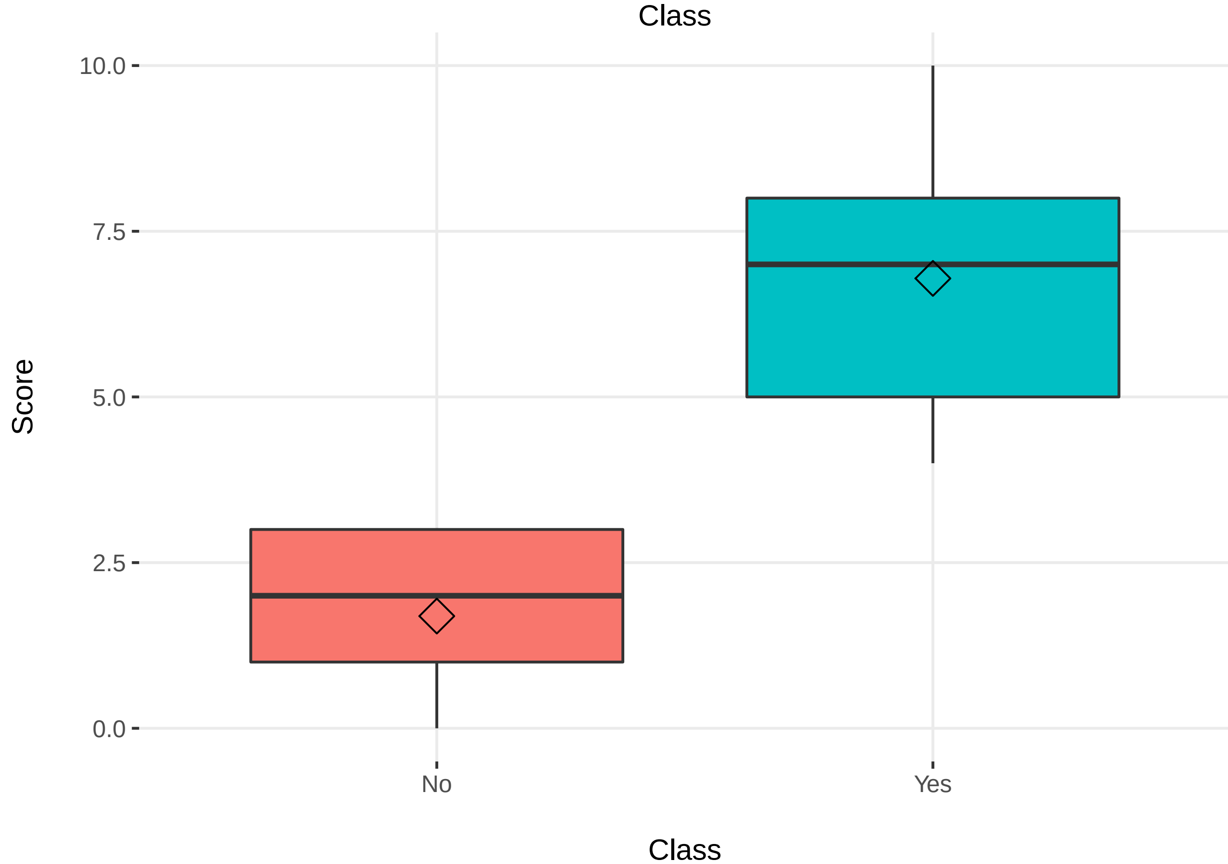

```
# Plot distribution between input and output variables.
# Reports if a variable is significant or not
# variables to analyse
f1 <- c("A9", "Score", "A5")
cross_plot(asd, input=f1, target = "Class")
```


Class (%)
100%
75%
50%
25%
0%
42.9
57.1
96.3
3.7
0
1
Class (count)
500
400
300
200
100
0
231
307
497
19
0
1
Yes
No
A9
Class (%)
0.00.00.00.0
100%
75%
50%
25%
0%
0.00.00.00.00.00.00.0
Score
A9
Class (count)
135
120
110
96
96
97
95
88
88
75
54
100
50
0
Score
Yes
No

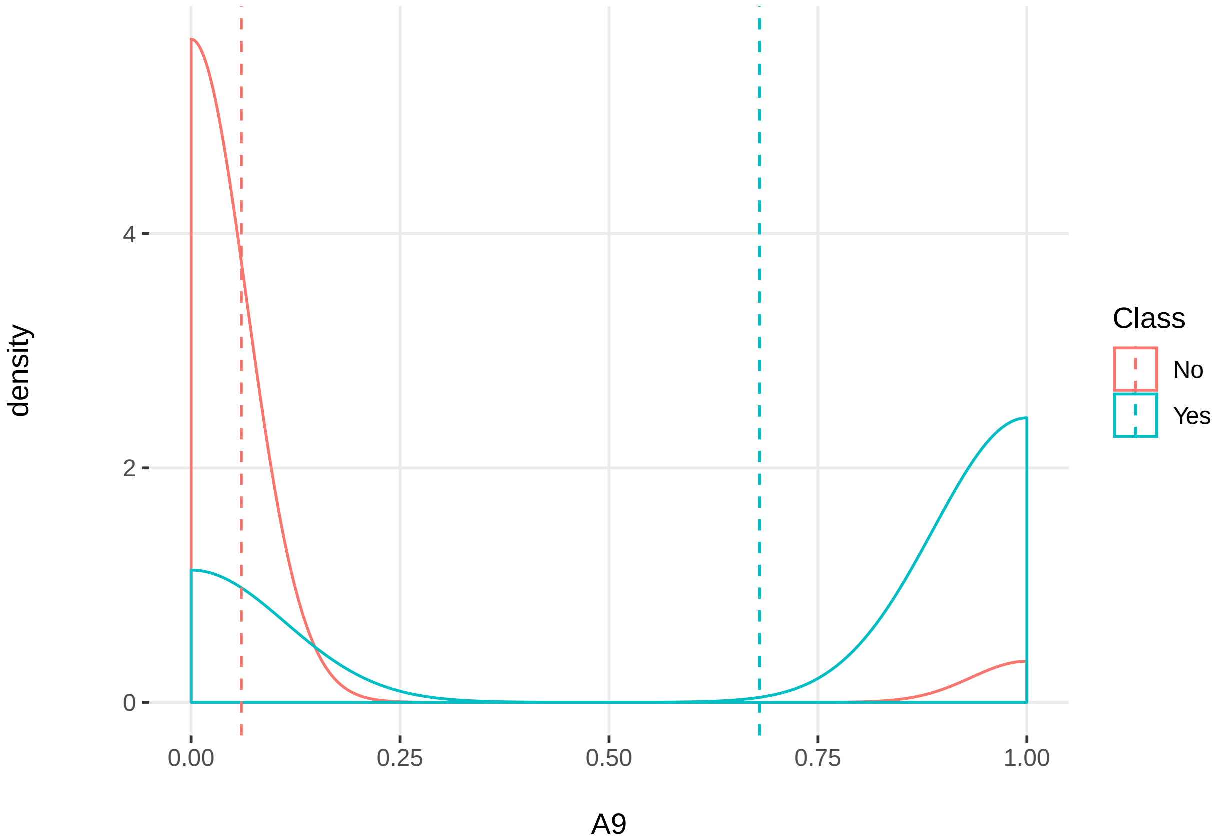


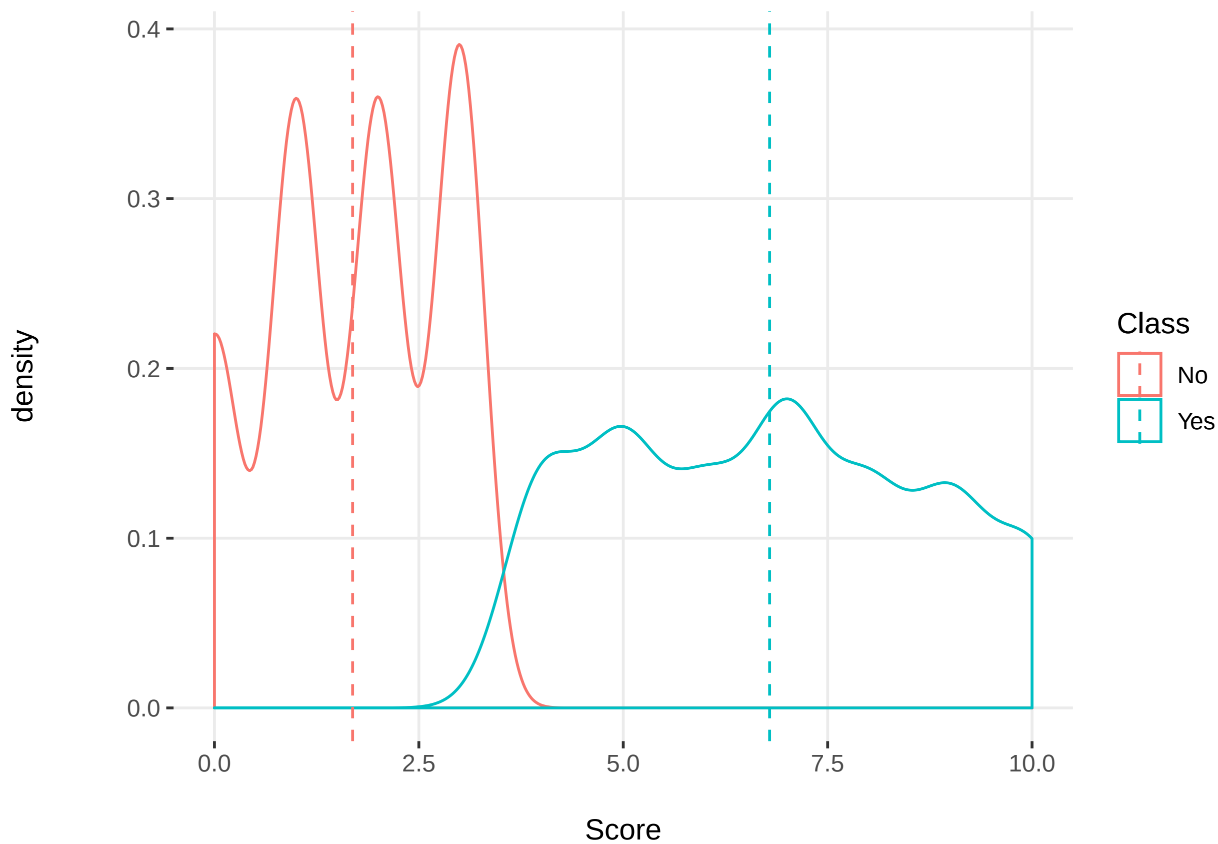

```r
# Report variable significance without Predictive Modelling based on Information Theory
plotar(asd, target = "Class", plot_type="boxplot")
```


30
20
Age_Mons
No
Yes
Class

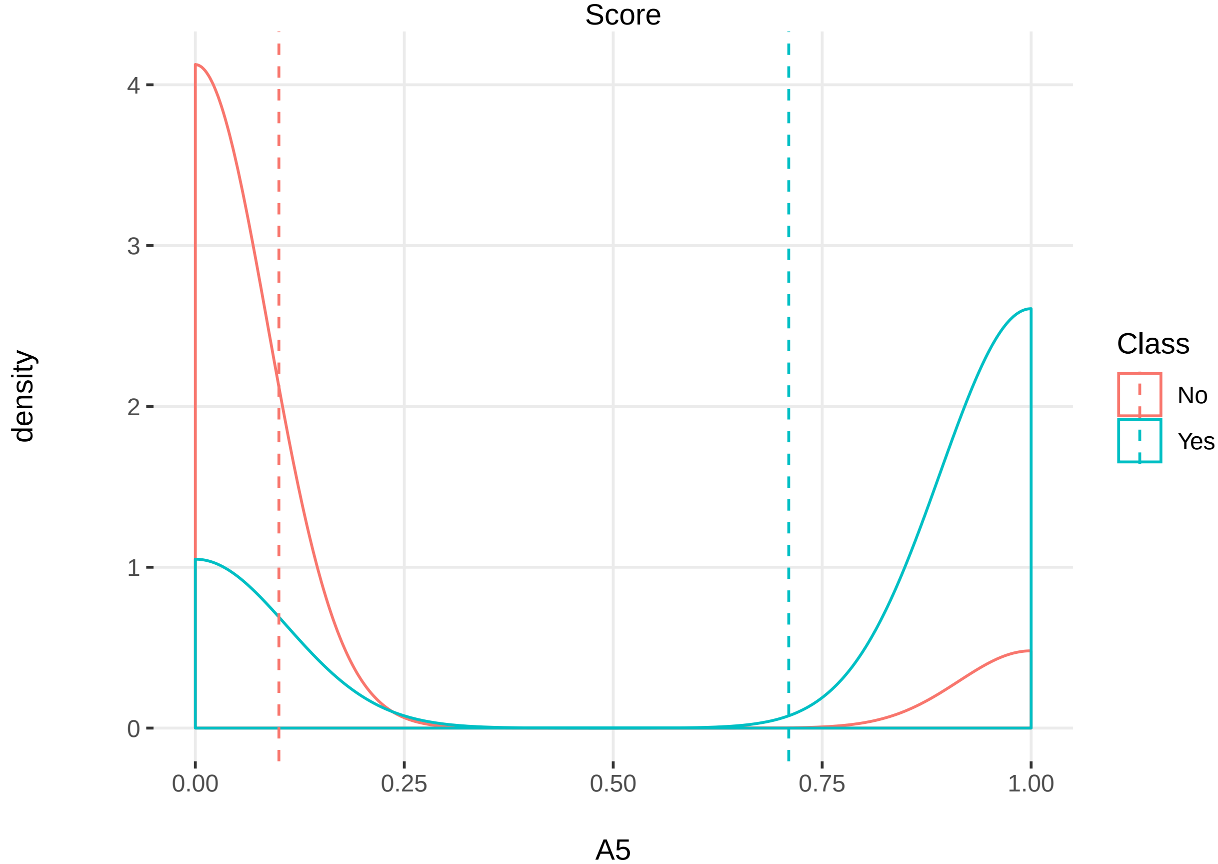

10.0
7.5
5.0
2.5
0.0
Score
No
Yes
Class

```r
plotar(asd, input=f1, target ="Class", plot_type="histdens")
```





density
0.4
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
Score
Class
No
Yes



density
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
A5
Class
No
Yes

Conclusion:

Findings in EDA are not final rather than suggestive in nature to investigate the correlations of the dependent variables and independent variables and might lead to answer the Problem. This process is part of Data Understanding for CRISP-DM framework that will assist us in the next stage which is Data Preprocessing.

Variables showing high correlation to the target Class based on Information Gain (IG)

Significant variables: A9, Score, A5, A6, A7, Score

Least significant variables:

Age_Mons, Sex, Ethnicity

Features to include

Important to include Jaundice feature to to build predictive model even if it has low information gain merit as this was identified a contributing factor to asd by medical practitioners.

Note: profiling for categorical variables were excluded as it's throwing errors during PDF compilation, will include in another file.