Statistical Analysis Report

Author: Aurora

Background

Autistic Spectrum Disorder (ASD) is a neurodevelopment condition which may have lots of costs of significant healthcare. So early diagnosis is necessary. However, the early ASD diagnosis need a long time. Thus, this analysis will seek to discover some variables, and build the model to determine the factors that predict probability of an autism diagnosis. It is helpful to the early diagnosis of ASD.

Data Source

Data is about the diagnosis of autism. There are 17 variables, X (unique identifier), A01 to A10 (these are all ten questions based on the screening method used), Age (age in years), Gender (male or female), Jaundic (whether the case was born with jaundice), Nationality (nationality of the autism case), Rel (who complete the test), Autism (diagnosed with autism)

Data Transformation and Cleaning (Description)

X : It was dropped for it is just the index of the data.

Nationality: It was transformed dummy variables.

Autism: It was transformed to binary.

Rel: It was transformed to dummy variables.

Rename these dummy variables:

Nationality_RWAfrican: NatAfr_RW

Nationality_RWAsian: NatAsi_RW

Nationality_RWEuropean: NatEur_RW

Nationality_RWLatin America: NatLatA_RW

Nationality_RWMiddle Eastern: NatMidE_RW

Nationality_RWNorth American: NatNorA_RW

Rel_RWHealth care professional: RelHCP_RW

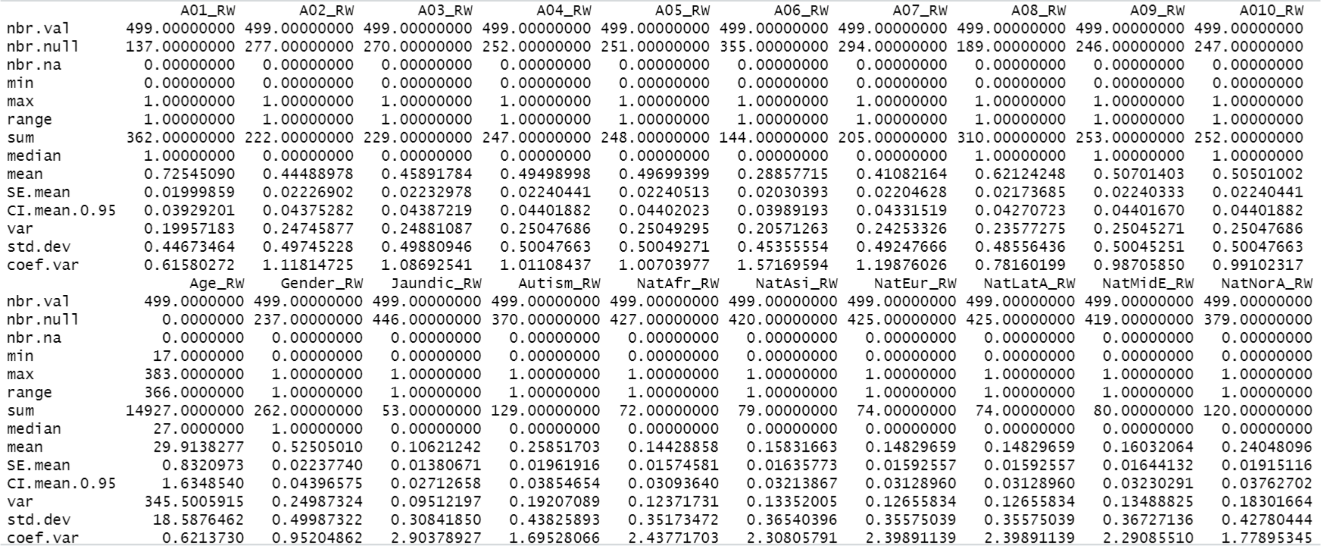
Rel_RWOthers: RelOth_RW

Rel_RWParent: RelPar_RW

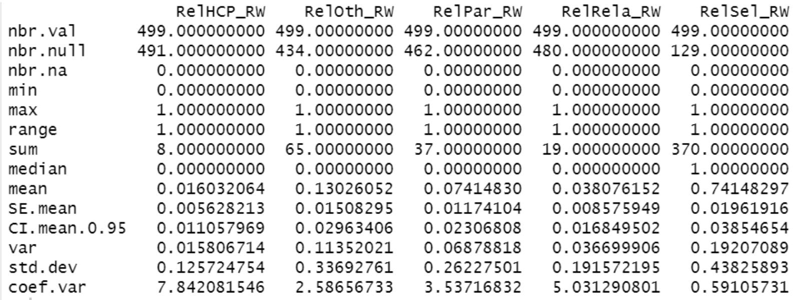
Rel_RWRelative: RelRela_RW

Rel_RWSelf: RelSel_RW

Descriptive Data Analysis

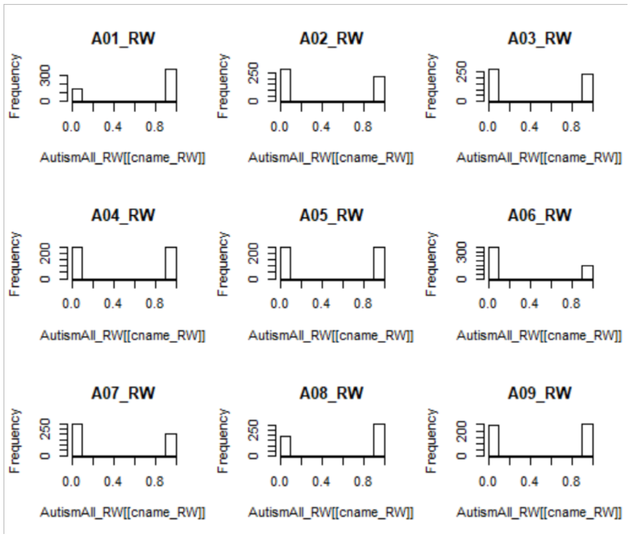


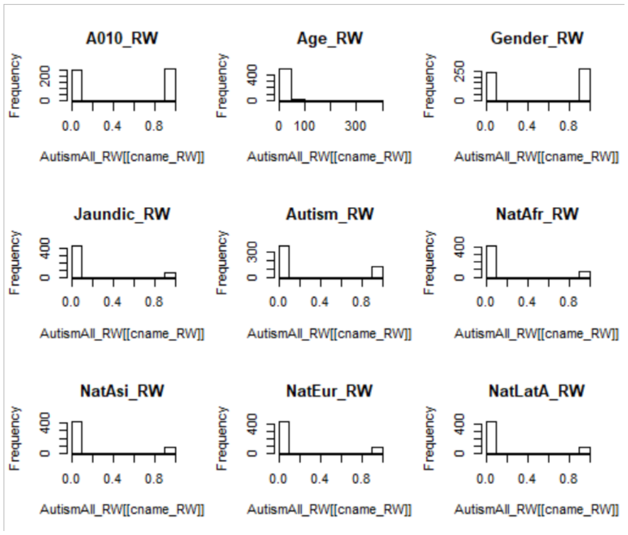
```
                  A01_RW       A02_RW       A03_RW       A04_RW       A05_RW       A06_RW       A07_RW       A08_RW       A09_RW      A010_RW
nbr.val      499.00000000 499.00000000 499.00000000 499.00000000 499.00000000 499.00000000 499.00000000 499.00000000 499.00000000 499.00000000
nbr.null     137.00000000 277.00000000 270.00000000 252.00000000 251.00000000 355.00000000 294.00000000 189.00000000 246.00000000 247.00000000
nbr.na         0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000
min            0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000
max            1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000
range          1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000
sum          362.00000000 222.00000000 229.00000000 247.00000000 248.00000000 144.00000000 205.00000000 310.00000000 253.00000000 252.00000000
median         1.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   1.00000000   1.00000000   1.00000000
mean           0.72545090   0.44488978   0.45891784   0.49498998   0.49699399   0.28857715   0.41082164   0.62124248   0.50701403   0.50501002
SE.mean        0.01999859   0.02226902   0.02232978   0.02240441   0.02240513   0.02030393   0.02204628   0.02173685   0.02240333   0.02240441
CI.mean.0.95   0.03929201   0.04375282   0.04387219   0.04401882   0.04402023   0.03989193   0.04331519   0.04270723   0.04401670   0.04401882
var            0.19957183   0.24745877   0.24881087   0.25047686   0.25049295   0.20571263   0.24253326   0.23577275   0.25045271   0.25047686
std.dev        0.44673464   0.49745228   0.49880946   0.50047663   0.50049271   0.45355554   0.49247666   0.48556436   0.50045251   0.50047663
coef.var       0.61580272   1.11814725   1.08692541   1.01108437   1.00703977   1.57169594   1.19876026   0.78160199   0.98705850   0.99102317
                     Age_RW    Gender_RW   Jaundic_RW    Autism_RW    NatAfr_RW    NatAsi_RW    NatEur_RW   NatLatA_RW   NatMidE_RW   NatNorA_RW
nbr.val        499.0000000 499.00000000 499.00000000 499.00000000 499.00000000 499.00000000 499.00000000 499.00000000 499.00000000 499.00000000
nbr.null         0.0000000 237.00000000 446.00000000 370.00000000 427.00000000 420.00000000 425.00000000 425.00000000 419.00000000 379.00000000
nbr.na           0.0000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000
min             17.0000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000
max            383.0000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000
range          366.0000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000   1.00000000
sum          14927.0000000 262.00000000  53.00000000 129.00000000  72.00000000  79.00000000  74.00000000  74.00000000  80.00000000 120.00000000
median          27.0000000   1.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000
mean            29.9138277   0.52505010   0.10621242   0.25851703   0.14428858   0.15831663   0.14829659   0.14829659   0.16032064   0.24048096
SE.mean          0.8320973   0.02237740   0.01380671   0.01961916   0.01574581   0.01635773   0.01592557   0.01592557   0.01644132   0.01915116
CI.mean.0.95     1.6348540   0.04396575   0.02712658   0.03854654   0.03093640   0.03213867   0.03128960   0.03128960   0.03230291   0.03762702
var            345.5005915   0.24987324   0.09512197   0.19207089   0.12371731   0.13352005   0.12655834   0.12655834   0.13488825   0.18301664
std.dev         18.5876462   0.49987322   0.30841850   0.43825893   0.35173472   0.36540396   0.35575039   0.35575039   0.36727136   0.42780444
coef.var         0.6213730   0.95204862   2.90378927   1.69528066   2.43771703   2.30805791   2.39891139   2.39891139   2.29085510   1.77895345
```

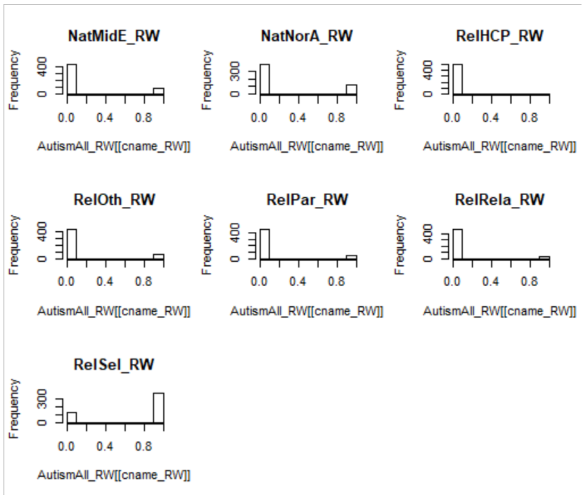


```
                 RelHCP_RW    RelOth_RW    RelPar_RW    RelRela_RW    RelSel_RW
nbr.val      499.000000000 499.00000000 499.00000000 499.000000000 499.00000000
nbr.null     491.000000000 434.00000000 462.00000000 480.000000000 129.00000000
nbr.na         0.000000000   0.00000000   0.00000000   0.000000000   0.00000000
min            0.000000000   0.00000000   0.00000000   0.000000000   0.00000000
max            1.000000000   1.00000000   1.00000000   1.000000000   1.00000000
range          1.000000000   1.00000000   1.00000000   1.000000000   1.00000000
sum            8.000000000  65.00000000  37.00000000  19.000000000 370.00000000
median         0.000000000   0.00000000   0.00000000   0.000000000   1.00000000
mean           0.016032064   0.13026052   0.07414830   0.038076152   0.74148297
SE.mean        0.005628213   0.01508295   0.01174104   0.008575949   0.01961916
CI.mean.0.95   0.011057969   0.02963406   0.02306808   0.016849502   0.03854654
var            0.015806714   0.11352021   0.06878818   0.036699906   0.19207089
std.dev        0.125724754   0.33692761   0.26227501   0.191572195   0.43825893
coef.var       7.842081546   2.58656733   3.53716832   5.031290801   0.59105731
```

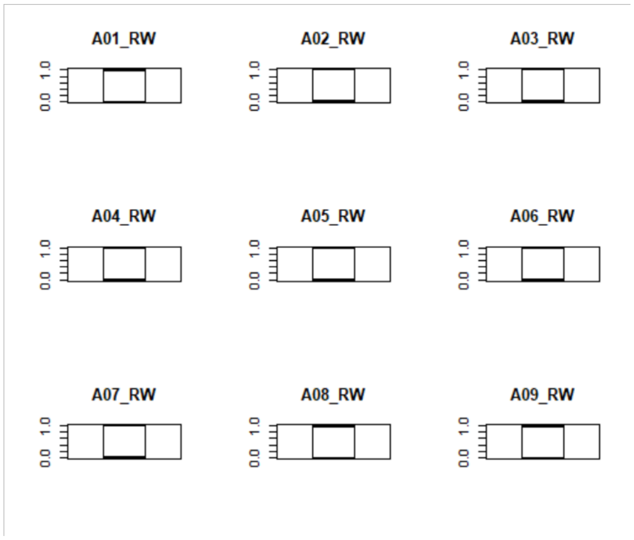
From the summary statistics we conclude that the transformation of data worked properly. And most of the data looks reasonable. But the max of Age is 383 which is unreasonable. I will deal with it later.

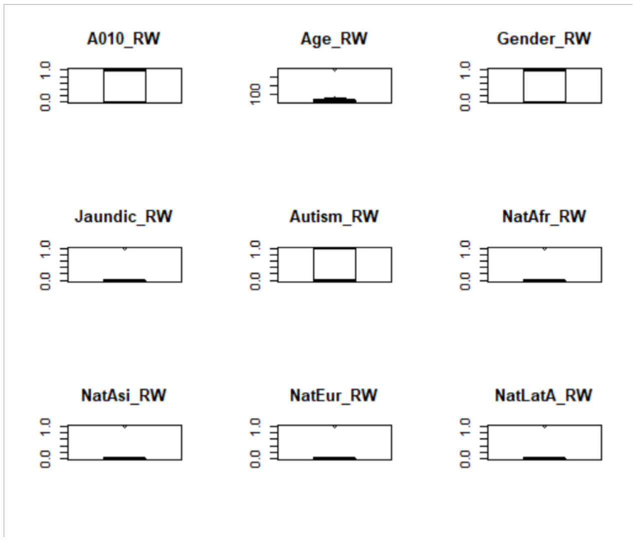


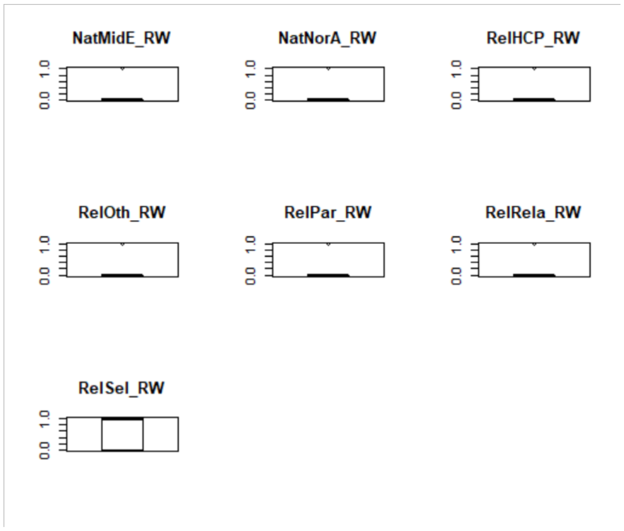



NatMidE_RW
NatNorA_RW
RelHCP_RW
Frequency
0
400
300
0.0
0.4
0.8
AutismAll_RW[[cname_RW]]
RelOth_RW
RelPar_RW
RelRela_RW
RelSel_RW

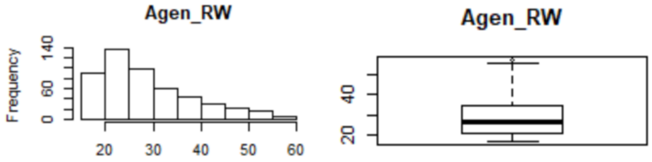
Outlier


A01_RW
A02_RW
A03_RW
A04_RW
A05_RW
A06_RW
A07_RW
A08_RW
A09_RW
0.0
1.0


A010_RW
Age_RW
Gender_RW
Jaundic_RW
Autism_RW
NatAfr_RW
NatAsi_RW
NatEur_RW
NatLatA_RW
0.0
1.0
100


NatMidE_RW
NatNorA_RW
RelHCP_RW
RelOth_RW
RelPar_RW
RelRela_RW
RelSel_RW
0.0
1.0

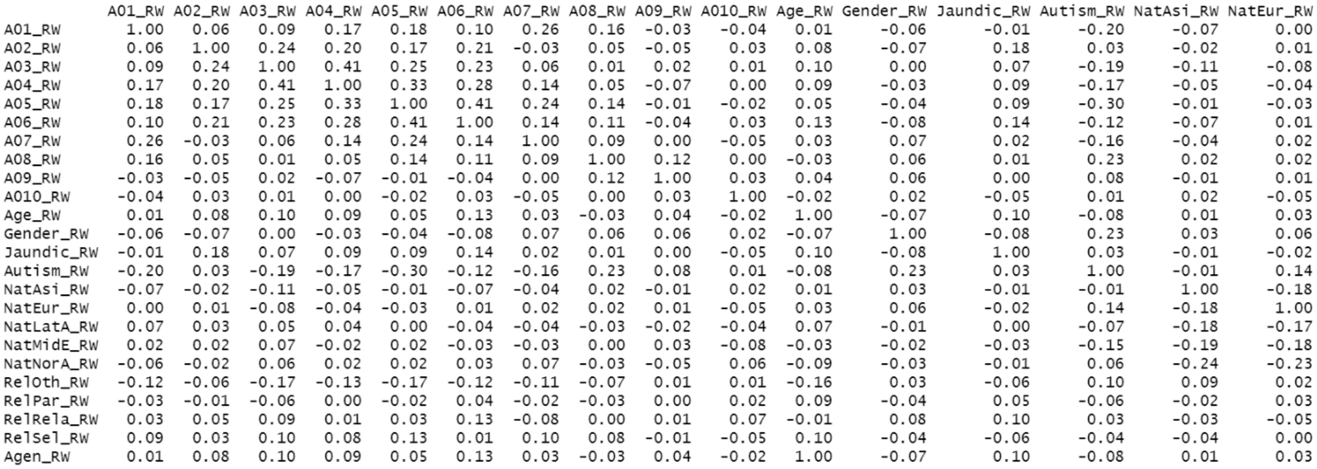
There seem to be outliers in Age. So I got the 99% quantile which was 56, and all the value which was higher than 56 equaled 57. After that, there is a new column called Agen as fllow:



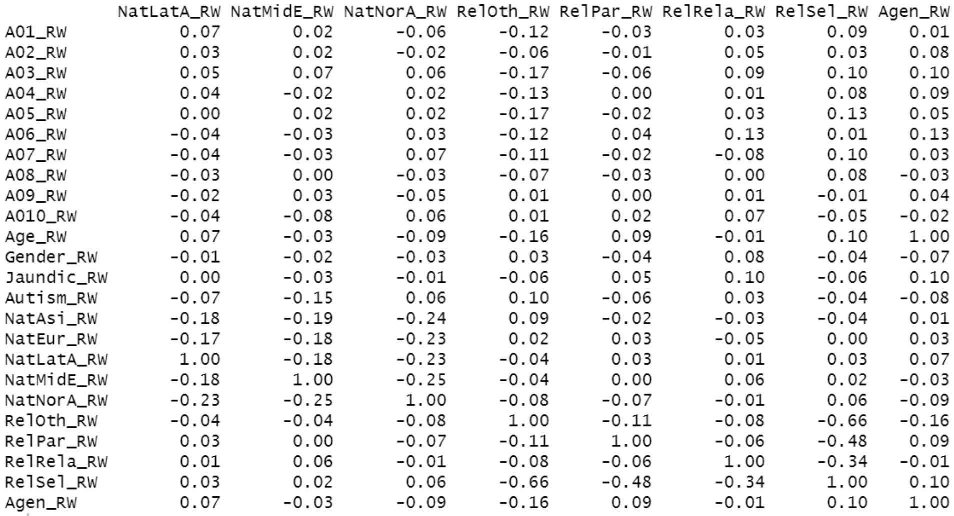
Exploratory Data Analysis

Correlations

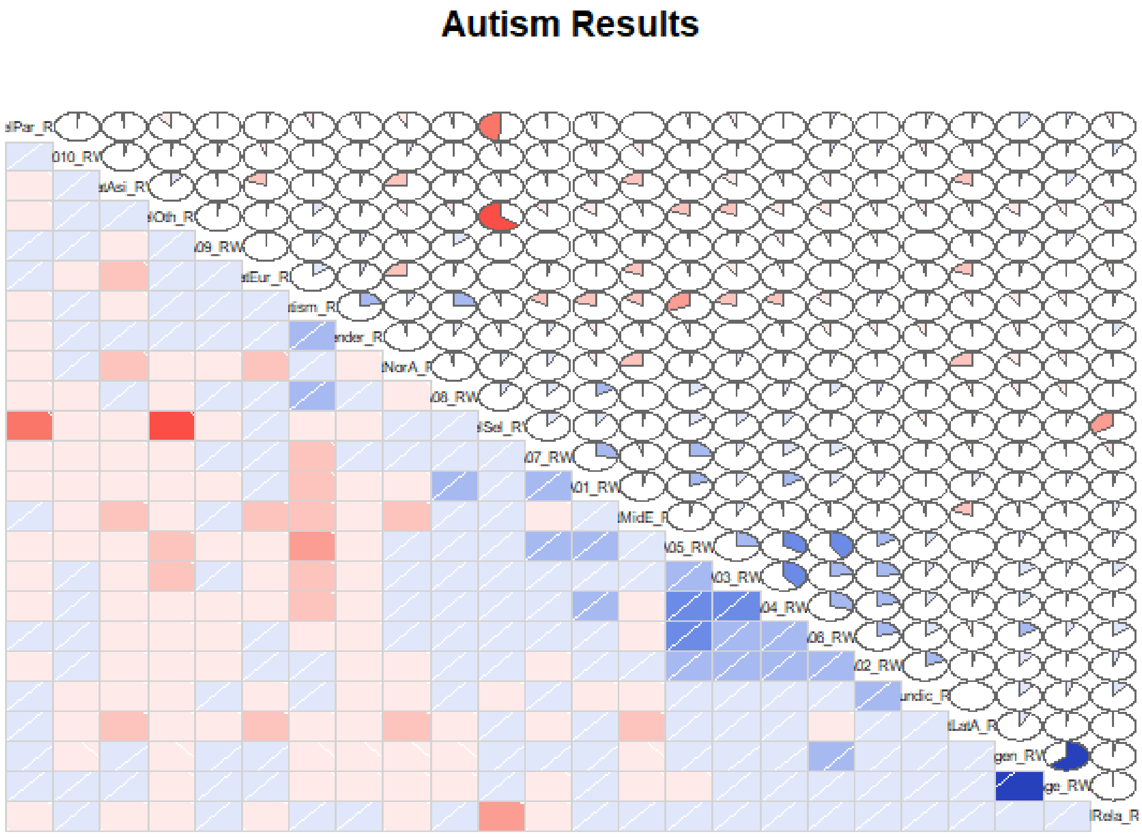
The column of 'NatAfri_RW', 'NatAsi_RW', 'NatEur_RW', 'NatLatA_RW', 'NatMidE_RW' and 'NatNorA_RW' are dummy variables. They are correlation, so I dropped 'NatAfri_RW'. And the column of 'RelHCP_RW', 'RelOth_RW', 'RelPar_RW', 'RelRela_RW' and 'RelSel_RW' are also dummy variables. They are correlation, so I dropped 'RelHCP_RW'.



```
           A01_RW A02_RW A03_RW A04_RW A05_RW A06_RW A07_RW A08_RW A09_RW A010_RW Age_RW Gender_RW Jaundic_RW Autism_RW NatAsi_RW NatEur_RW
A01_RW       1.00   0.06   0.09   0.17   0.18   0.10   0.26   0.16  -0.03   -0.04   0.01     -0.06      -0.01     -0.20     -0.07      0.00
A02_RW       0.06   1.00   0.24   0.20   0.17   0.21  -0.03   0.05  -0.05    0.03   0.08     -0.07       0.18      0.03     -0.02      0.01
A03_RW       0.09   0.24   1.00   0.41   0.25   0.23   0.06   0.01   0.02    0.01   0.10      0.00       0.07     -0.19     -0.11     -0.08
A04_RW       0.17   0.20   0.41   1.00   0.33   0.28   0.14   0.05  -0.07    0.00   0.09     -0.03       0.09     -0.17     -0.05     -0.04
A05_RW       0.18   0.17   0.25   0.33   1.00   0.41   0.24   0.14  -0.01   -0.02   0.05     -0.04       0.09     -0.30     -0.01     -0.03
A06_RW       0.10   0.21   0.23   0.28   0.41   1.00   0.14   0.11  -0.04    0.03   0.13     -0.08       0.14     -0.12     -0.07      0.01
A07_RW       0.26  -0.03   0.06   0.14   0.24   0.14   1.00   0.09   0.00   -0.05   0.03      0.07       0.02     -0.16     -0.04      0.02
A08_RW       0.16   0.05   0.01   0.05   0.14   0.11   0.09   1.00   0.12    0.00  -0.03      0.06       0.01      0.23      0.02      0.02
A09_RW      -0.03  -0.05   0.02  -0.07  -0.01  -0.04   0.00   0.12   1.00    0.03   0.04      0.06       0.00      0.08     -0.01      0.01
A010_RW     -0.04   0.03   0.01   0.00  -0.02   0.03  -0.05   0.00   0.03    1.00  -0.02      0.02      -0.05      0.01      0.02     -0.05
Age_RW       0.01   0.08   0.10   0.09   0.05   0.13   0.03  -0.03   0.04   -0.02   1.00     -0.07       0.10     -0.08      0.01      0.03
Gender_RW   -0.06  -0.07   0.00  -0.03  -0.04  -0.08   0.07   0.06   0.06    0.02  -0.07      1.00      -0.08      0.23      0.03      0.06
Jaundic_RW  -0.01   0.18   0.07   0.09   0.09   0.14   0.02   0.01   0.00   -0.05   0.10     -0.08       1.00      0.03     -0.01     -0.02
Autism_RW   -0.20   0.03  -0.19  -0.17  -0.30  -0.12  -0.16   0.23   0.08    0.01  -0.08      0.23       0.03      1.00     -0.01      0.14
NatAsi_RW   -0.07  -0.02  -0.11  -0.05  -0.01  -0.07  -0.04   0.02  -0.01    0.02   0.01      0.03      -0.01     -0.01      1.00     -0.18
NatEur_RW    0.00   0.01  -0.08  -0.04  -0.03   0.01   0.02   0.02   0.01   -0.05   0.03      0.06      -0.02      0.14     -0.18      1.00
NatLatA_RW   0.07   0.03   0.05   0.04   0.00  -0.04  -0.04  -0.03  -0.02   -0.04   0.07     -0.01       0.00     -0.07     -0.18     -0.17
NatMidE_RW   0.02   0.02   0.07  -0.02   0.02  -0.03  -0.03   0.00   0.03   -0.08  -0.03     -0.02      -0.03     -0.15     -0.19     -0.18
NatNorA_RW  -0.06  -0.02   0.06   0.02   0.02   0.03   0.07  -0.03  -0.05    0.06  -0.09     -0.03      -0.01      0.06     -0.24     -0.23
RelOth_RW   -0.12  -0.06  -0.17  -0.13  -0.17  -0.12  -0.11  -0.07   0.01    0.01  -0.16      0.03      -0.06      0.10      0.09      0.02
RelPar_RW   -0.03  -0.01  -0.06   0.00  -0.02   0.04  -0.02  -0.03   0.00    0.02   0.09     -0.04       0.05     -0.06     -0.02      0.03
RelRela_RW   0.03   0.05   0.09   0.01   0.03   0.13  -0.08   0.00   0.01    0.07  -0.01      0.08       0.10      0.03     -0.03     -0.05
RelSel_RW    0.09   0.03   0.10   0.08   0.13   0.01   0.10   0.08  -0.01   -0.05   0.10     -0.04      -0.06     -0.04     -0.04      0.00
Agen_RW      0.01   0.08   0.10   0.09   0.05   0.13   0.03  -0.03   0.04   -0.02   1.00     -0.07       0.10     -0.08      0.01      0.03
```



```
           NatLatA_RW NatMidE_RW NatNorA_RW RelOth_RW RelPar_RW RelRela_RW RelSel_RW Agen_RW
A01_RW           0.07       0.02      -0.06     -0.12     -0.03       0.03      0.09    0.01
A02_RW           0.03       0.02      -0.02     -0.06     -0.01       0.05      0.03    0.08
A03_RW           0.05       0.07       0.06     -0.17     -0.06       0.09      0.10    0.10
A04_RW           0.04      -0.02       0.02     -0.13      0.00       0.01      0.08    0.09
A05_RW           0.00       0.02       0.02     -0.17     -0.02       0.03      0.13    0.05
A06_RW          -0.04      -0.03       0.03     -0.12      0.04       0.13      0.01    0.13
A07_RW          -0.04      -0.03       0.07     -0.11     -0.02      -0.08      0.10    0.03
A08_RW          -0.03       0.00      -0.03     -0.07     -0.03       0.00      0.08   -0.03
A09_RW          -0.02       0.03      -0.05      0.01      0.00       0.01     -0.01    0.04
A010_RW         -0.04      -0.08       0.06      0.01      0.02       0.07     -0.05   -0.02
Age_RW           0.07      -0.03      -0.09     -0.16      0.09      -0.01      0.10    1.00
Gender_RW       -0.01      -0.02      -0.03      0.03     -0.04       0.08     -0.04   -0.07
Jaundic_RW       0.00      -0.03      -0.01     -0.06      0.05       0.10     -0.06    0.10
Autism_RW       -0.07      -0.15       0.06      0.10     -0.06       0.03     -0.04   -0.08
NatAsi_RW       -0.18      -0.19      -0.24      0.09     -0.02      -0.03     -0.04    0.01
NatEur_RW       -0.17      -0.18      -0.23      0.02      0.03      -0.05      0.00    0.03
NatLatA_RW       1.00      -0.18      -0.23     -0.04      0.03       0.01      0.03    0.07
NatMidE_RW      -0.18       1.00      -0.25     -0.04      0.00       0.06      0.02   -0.03
NatNorA_RW      -0.23      -0.25       1.00     -0.08     -0.07      -0.01      0.06   -0.09
RelOth_RW       -0.04      -0.04      -0.08      1.00     -0.11      -0.08     -0.66   -0.16
RelPar_RW        0.03       0.00      -0.07     -0.11      1.00      -0.06     -0.48    0.09
RelRela_RW       0.01       0.06      -0.01     -0.08     -0.06       1.00     -0.34   -0.01
RelSel_RW        0.03       0.02       0.06     -0.66     -0.48      -0.34      1.00    0.10
Agen_RW          0.07      -0.03      -0.09     -0.16      0.09      -0.01      0.10    1.00
```



There are some correlations:

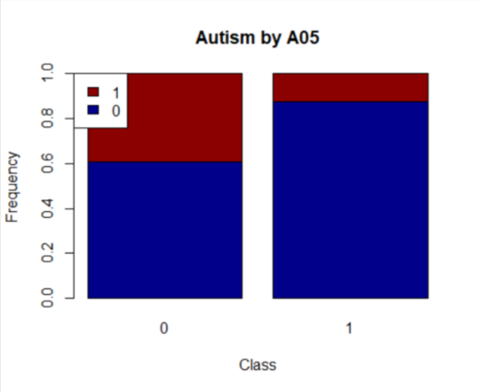
A05_RW and A06_RW have positive correlation (0.41). RelSel_RW and RelOth_RW have negative correlation (-0.66). RelSel_RW and RelPar_RW have negative correlation (-0.48)

Two most significant predictors of Autism is A05_RW (-0.3) and Gender_RW(0.23) .

1. A05_RW and Autism_RW

```
     0   1
  0 153  98
  1 217  31

         0        1
  0 60.95618 39.04382
  1 87.50000 12.50000

Number of cases in table: 499
Number of factors: 2
Test for independence of all factors:
        Chisq = 45.85, df = 1, p-value = 0.00000000001275

Pearson's Chi-squared test with Yates' continuity correction

data:  AutismAll_RW$A05_RW and AutismAll_RW$Autism_RW
```

```
X-squared = 44.478, df = 1, p-value = 0.00000000002572
```

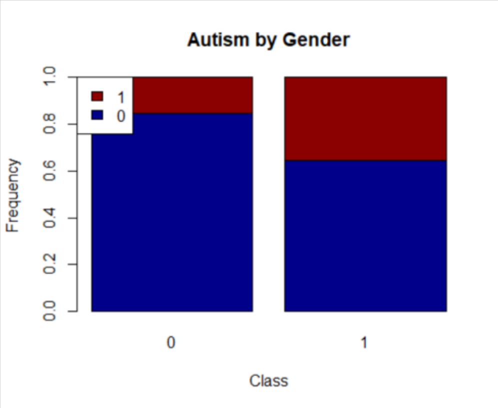


There is 60.96%(153 persons) of people whose the answer of question 5 is 0 is not Autism.
There is 87.5%(217 persons) of people whose the answer of question 5 is 1 is not Autism
There is 39%(98 persons) of people whose the answer of question 5 is 0 is Autism
There is 12.5%(31 persons) of people whose the answer of question 5 is 1 is Autism

Because of the p value (<0.05) of Pearson's Chi-squared test, the variable of A05_RW and Autism_RW are not correlation.

2. A05_RW and Autism_RW

```
     0   1
  0 201  36
  1 169  93

         0        1
  0 84.81013 15.18987
  1 64.50382 35.49618

Number of cases in table: 499 
Number of factors: 2 
Test for independence of all factors:
        Chisq = 26.768, df = 1, p-value = 0.0000002294

Pearson's Chi-squared test with Yates' continuity correction

data:  AutismAll_RW$Gender_RW and AutismAll_RW$Autism_RW
X-squared = 25.719, df = 1, p-value = 0.0000003948
```



There is 84.81%(201 persons) of female is not Autism.
There is 64.50%(169 persons) of male is not Autism.
There is 15.20%(36 persons) of female is Autism.
There is 35.50%(93 persons) of male is Autism.

Because of the p value (<0.05) of Pearson's Chi-squared test, the variable of Gender_RW and Autism_RW are not correlation.

Models

Model 1: Baseline Model with all Variables included

1. Fisher's Scoring Interation is 6 which means the model converge until the 6th interation.
2. AIC is 428.68.
3. The residuals look approximately symmetrical. The residual deviance is 382.68 which is less than null deviance(570.36). That is normal.
4. Seven variables (A01_RW, A02_RW, A03_RW, A05_RW, A08_RW, Gender_RW, NatMidE_RW) have Z-values less than 0.05. These variables look significant.
5. A010_RW is negatively correlated with Autism_RW instead of positively.
   RelPar_RW is positively correlated with Autism_RW instead of negatively.
   RelSel_RW is positively correlated with Autism_RW instead of negatively.

```
Call:
glm(formula = Autism_RW ~ A01_RW + A02_RW + A03_RW + A04_RW + 
    A05_RW + A06_RW + A07_RW + A08_RW + A09_RW + A010_RW + Agen_RW + 
    Gender_RW + Jaundic_RW + NatAsi_RW + NatEur_RW + NatLatA_RW + 
    NatMidE_RW + NatNorA_RW + RelOth_RW + RelPar_RW + RelRela_RW + 
    RelSel_RW, family = "binomial", data = AutismAll_RW, na.action = na.omit)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.8089  -0.6091  -0.2939   0.4118   3.0296  

Coefficients:
            Estimate Std. Error z value       Pr(>|z|)    
(Intercept) -2.62082    1.51653  -1.728        0.08396 .  
A01_RW      -0.85366    0.30072  -2.839        0.00453 ** 
A02_RW       0.87477    0.28621   3.056        0.00224 ** 
A03_RW      -0.67751    0.29945  -2.263        0.02366 *  
A04_RW      -0.35898    0.30044  -1.195        0.23214    
A05_RW      -1.69939    0.32001  -5.311 0.000000109318 ***
A06_RW      -0.10241    0.34525  -0.297        0.76675    
A07_RW      -0.56866    0.29317  -1.940        0.05241 .  
A08_RW       2.01186    0.32400   6.210 0.000000000531 ***
A09_RW       0.32237    0.26296   1.226        0.22024    
A010_RW     -0.18601    0.26236  -0.709        0.47832    
Agen_RW     -0.01061    0.01472  -0.720        0.47124    
Gender_RW    1.41089    0.28785   4.901 0.000000951323 ***
Jaundic_RW   0.63858    0.43675   1.462        0.14370    
NatAsi_RW   -0.33433    0.47121  -0.710        0.47800    
NatEur_RW    0.65441    0.45435   1.440        0.14977    
NatLatA_RW  -0.42810    0.49473  -0.865        0.38686    
NatMidE_RW  -1.34720    0.55379  -2.433        0.01499 *  
```

```
NatNorA_RW   0.52652    0.42579   1.237        0.21625
RelOth_RW    1.21105    1.40813   0.860        0.38976
RelPar_RW    0.19851    1.46706   0.135        0.89236
RelRela_RW   1.69256    1.48702   1.138        0.25503
RelSel_RW    1.07556    1.36695   0.787        0.43138
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 570.36  on 498  degrees of freedom
Residual deviance: 382.68  on 476  degrees of freedom
AIC: 428.68

Number of Fisher Scoring iterations: 6
```

Model 2: Stepwise Selection Model

1. Fisher’s Scoring Interation is 5 which means the model converge until 5th interation.
2. AIC is 418.09.
3. The residuals look approximately symmetrical. The residual deviance is 388.09 which is less than null deviance(570.36). That is normal.
4. Eleven variables (A01_RW, A02_RW, A03_RW, A05_RW , A07_RW , A08_RW ,Gender_RW, NatEur_RW, NatMidE_RW, NatNorA_RW, RelRela_RW) and intercept have Z-values less than 0.05. These variables look significant.
5. RelSel_RW is positively correlated with Autism_RW instead of negatively.

```
Call:
glm(formula = Autism_RW ~ A01_RW + A02_RW + A03_RW + A05_RW + 
    A07_RW + A08_RW + Gender_RW + Jaundic_RW + NatEur_RW + NatMidE_RW + 
    NatNorA_RW + RelOth_RW + RelRela_RW + RelSel_RW, family = "binomial", 
    data = AutismAll_RW, na.action = na.omit)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.8213  -0.5981  -0.3236   0.4200   2.8942  

Coefficients:
            Estimate Std. Error z value       Pr(>|z|)    
(Intercept)  -2.9957     0.6448  -4.646 0.000003387219 ***
A01_RW       -0.8665     0.2950  -2.938        0.00331 ** 
A02_RW        0.7886     0.2802   2.815        0.00488 ** 
A03_RW       -0.8142     0.2821  -2.886        0.00391 ** 
A05_RW       -1.7764     0.2991  -5.939 0.000000002870 ***
A07_RW       -0.5860     0.2879  -2.036        0.04179 *  
A08_RW        1.9879     0.3142   6.326 0.000000000252 ***
Gender_RW     1.3827     0.2805   4.929 0.000000827191 ***
Jaundic_RW    0.5996     0.4178   1.435        0.15132    
NatEur_RW     0.9152     0.3560   2.571        0.01015 *  
NatMidE_RW   -0.9203     0.4614  -1.995        0.04608 *  
NatNorA_RW    0.8138     0.3190   2.551        0.01074 *  
RelOth_RW     1.1449     0.6045   1.894        0.05824 .  
RelRela_RW    1.5419     0.7730   1.995        0.04607 *  
RelSel_RW     0.9399     0.5187   1.812        0.06996 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 570.36  on 498  degrees of freedom
Residual deviance: 388.09  on 484  degrees of freedom
AIC: 418.09

Number of Fisher Scoring iterations: 5
```

Model 3: Additional Model -1

Based on the Stepwise Model, I select the variables which are significant according to Z-value. So this Additional Model-1 is on the variables of A01_RW, A02_RW, A03_RW, A05_RW, A07_RW, A08_RW, Gender_RW, NatEur_RW, NatMidE_RW, NatNorA_RW and RelRela_RW

1. Fisher's Scoring Interation is 5 which means the model converge until 5th interation.
2. AIC is 417.12.
3. The residuals look approximately symmetrical. The residual deviance is 395.12 which is less than null deviance(570.36). That is normal.
4. Except the NatMidE_RW, all the other variables and intercept have Z-values less than 0.05. These variables look significant.
5. All variable co-efficients show the correlation correctly.

```
Call:
glm(formula = Autism_RW ~ A01_RW + A02_RW + A03_RW + A05_RW + 
    A07_RW + A08_RW + Gender_RW + NatEur_RW + NatMidE_RW + NatNorA_RW, 
    family = "binomial", data = AutismAll_RW, na.action = na.omit)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.1064  -0.6329  -0.3292   0.4609   2.8542  

Coefficients:
            Estimate Std. Error z value       Pr(>|z|)    
(Intercept)  -2.0073     0.3952  -5.080 0.000000378237 ***
A01_RW       -0.8805     0.2896  -3.041        0.00236 ** 
A02_RW        0.8668     0.2751   3.151        0.00163 ** 
A03_RW       -0.7973     0.2764  -2.885        0.00392 ** 
A05_RW       -1.7147     0.2924  -5.864 0.000000004524 ***
A07_RW       -0.6337     0.2819  -2.248        0.02458 *  
A08_RW        1.9587     0.3090   6.339 0.000000000232 ***
Gender_RW     1.3546     0.2741   4.942 0.000000771432 ***
NatEur_RW     0.8989     0.3478   2.584        0.00976 ** 
NatMidE_RW   -0.8480     0.4490  -1.889        0.05896 .  
NatNorA_RW    0.7972     0.3160   2.523        0.01164 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 570.36  on 498  degrees of freedom
Residual deviance: 395.12  on 488  degrees of freedom
AIC: 417.12

Number of Fisher Scoring iterations: 5
```

Model 4: Additional Model-2

Based on the Additional Model-1, I drop the variable of NatMidE_RW, because it is not significant in Additional Model-1.

1. Fisher's Scoring Interation is 5 which means the model converge until 5th interation.
2. AIC is 419.03.
3. The residuals look approximately symmetrical. The residual deviance is 399.03 which is less than null deviance(570.36). That is normal.
4. All the other variables and intercept have Z-values less than 0.05. These variables look significant.
5. All variable co-efficients show the correlation correctly.

```
Call:
glm(formula = Autism_RW ~ A01_RW + A02_RW + A03_RW + A05_RW + 
    A07_RW + A08_RW + Gender_RW + NatEur_RW + NatNorA_RW, family = "binomial", 
    data = AutismAll_RW, na.action = na.omit)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.1197  -0.6271  -0.3230   0.4599   2.7394  

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)    
(Intercept)  -2.1566     0.3885  -5.551 0.00000002832 ***
A01_RW       -0.8881     0.2875  -3.089       0.00201 ** 
A02_RW        0.8459     0.2731   3.098       0.00195 ** 
A03_RW       -0.8279     0.2757  -3.004       0.00267 ** 
A05_RW       -1.7270     0.2910  -5.936 0.00000000293 ***
A07_RW       -0.6360     0.2819  -2.256       0.02408 *  
A08_RW        1.9572     0.3077   6.361 0.00000000020 ***
Gender_RW     1.3511     0.2725   4.958 0.00000071365 ***
NatEur_RW     1.0829     0.3378   3.206       0.00135 ** 
NatNorA_RW    0.9831     0.3050   3.224       0.00126 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 570.36  on 498  degrees of freedom
Residual deviance: 399.03  on 489  degrees of freedom
AIC: 419.03

Number of Fisher Scoring iterations: 5
```

Model Evaluation

Additional Model 1

1. **Confusion Matrix**

```
      Yhat = 1 Yhat = 0
Y = 1       84       45
Y = 0       54      316
```

```
Stp1AutGlm_RW$Accurary_RW  0.8016032
Stp1AutGlm_RW$Specif_RW    0.8540541
Stp1AutGlm_RW$Recall_RW    0.6511628
Stp1AutGlm_RW$Precis_RW    0.6086957
```

According to the confusion matrix of additional model 1, there are 84 records which are predicted to Autism and they are actually Autism. There are 45 records which are predicted to non-autism and they are actually Autism. There are 54 records which are predicted to Autism and they are actually not Autism. There are 316 records which are predicted to non-autism and they are actually not Autism.

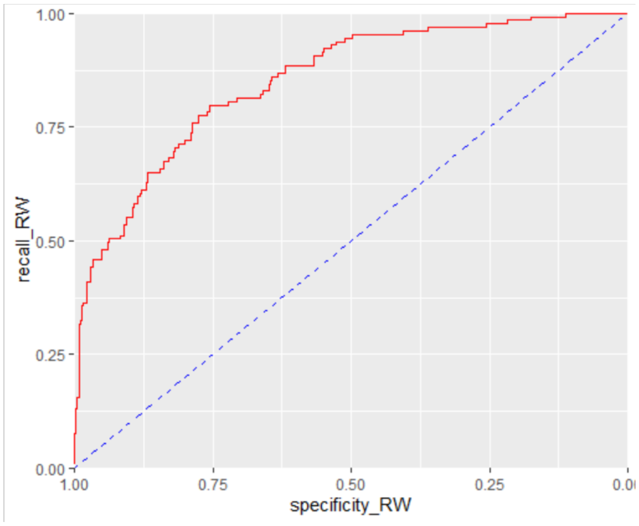
Accurary is 80.16% which means the proportion of cases correctly classified by this model.

Specificity is 85.40% which means percent of non-autism classified as non-autism.

Recall is 65.12% which means percent of Autism classified as Autism.

Precision is 61.70% which means percent of predicted Autism that are Autism.

2. **ROC Curve**



The ROC curve means the relation between recall and specificity. The curve is more away more better, because the curve is away from the bule line when recall and specificity both high.

3. **AUC**

```
AUC_RW 0.8510371
```

AUC means the area under curve, this value is more higher more better.

Additional Model 2

1. **Confusion Matrix**

```
        Yhat = 1 Yhat = 0
Y = 1        81       48
Y = 0        55      315

Stp2AutGlm_RW$Accurary_RW  0.7935872
Stp2AutGlm_RW$Specif_RW    0.8513514
Stp2AutGlm_RW$Recall_RW    0.627907
Stp2AutGlm_RW$Precis_RW    0.5955882
```

According to the confusion matrix of additional model 2, there are 81 records which are predicted to Autism and they are actually Autism. There are 48 records which are predicted to non-autism and they are actually Autism. There are 55 records which are predicted to Autism and they are actually not Autism. There are 315 records which are predicted to non-autism and they are actually not Autism.

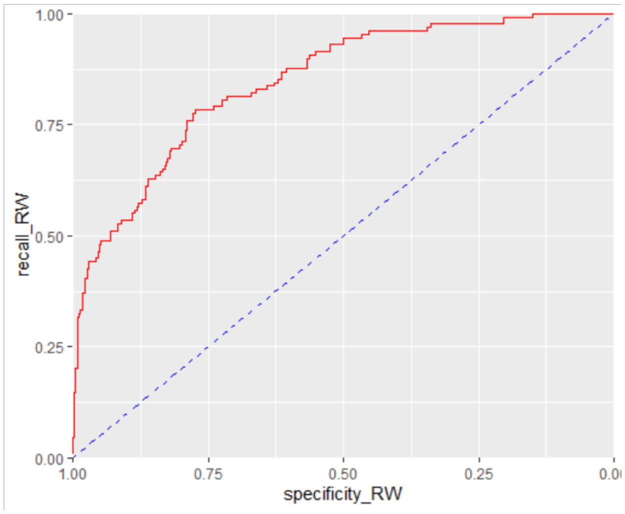
Accurary is 79.36% which means the proportion of cases correctly classified by this model.

Specificity is 85.14% which means percent of non-autism classified as non-autism.

Recall is 62.79% which means percent of Autism classified as Autism.

Precision is 59.56% which means percent of predicted Autism that are Autism.

2. **ROC Curve**



The ROC curve means the relation between recall and specificity. The curve is more away more better, because the curve is away from the bule line when recall and specificity both high.

3. **AUC**

```
AUC_RW   0.8474335
```

AUC means the area under curve, this value is more higher more better.

Final Model, Recommendation and Interpretation

Based on the above, I choose the additional model 1, for the AIC of this model is the lowest one of these four model. According to confusion matrix, the accuracy, specificity and precision of additional model 1 is higher than additional model 2. The ROC curve of additional model 1 looks better which means the red curve seems higher. The value of AUC of additional model 1 is higher. So the additional model 1 is the best of these four model.

Autism = (-0.8805) *A01_RW + 0.8668*A02_RW + (-0.7973) *A03_RW + (-1.7147) *A05_RW +

(-0.6337) *A07_RW + 1.9587*A08_RW + 1.3546*Gender_RW + 0.8989*NatEur_RW +

(-0.8480) *NatMidE_RW + 0.7972*NatNorA_RW+(-2.0073)