

Autism Screening Data Analysis Project

2025-11-04

About data analysis

The aim of this study is to explore how demographic factors (gender, ethnicity, and family autism history) and individual AQ-10 item responses relate to autism classification in adults.

Research questions:

- Which of the ten AQ-10 items (A1–A10) are most strongly associated with an autism diagnosis?
- Is there a difference in the distribution of autism diagnoses between genders?
- Do males and females with autism differ in their average AQ-10 total scores?
- Does the prevalence of autism vary across different ethnic groups?
- Is having a family history of autism associated with higher AQ-10 scores?

Loading data and initial data exploration

```
df <- read.csv("./dataset/Autism Screening.csv")  
  
head(df)  
  
##   A1_Score A2_Score A3_Score A4_Score A5_Score A6_Score A7_Score A8_Score  
## 1       1       1       1       1       0       0       1       1  
## 2       1       1       0       1       0       0       0       1  
## 3       1       1       0       1       1       0       1       1  
## 4       1       1       0       1       0       0       1       1  
## 5       1       0       0       0       0       0       0       1  
## 6       1       1       1       1       1       0       1       1  
##   A9_Score A10_Score age gender      ethnicity jundice austim    contry_of_res  
## 1       0       0   26     f White-European      no      no 'United States'  
## 2       0       1   24     m      Latino      no     yes      Brazil  
## 3       1       1   27     m      Latino      yes     yes      Spain  
## 4       0       1   35     f White-European      no     yes 'United States'  
## 5       0       0   40     f           ?      no      no      Egypt  
## 6       1       1   36     m      Others      yes      no 'United States'  
##   used_app_before result      age_desc relation Class.ASD  
## 1           no      6 '18 and more'     Self      NO  
## 2           no      5 '18 and more'     Self      NO  
## 3           no      8 '18 and more'   Parent      YES  
## 4           no      6 '18 and more'     Self      NO  
## 5           no      2 '18 and more'     ?      NO  
## 6           no      9 '18 and more'     Self      YES
```

```

names(df)

## [1] "A1_Score"      "A2_Score"      "A3_Score"      "A4_Score"
## [5] "A5_Score"      "A6_Score"      "A7_Score"      "A8_Score"
## [9] "A9_Score"      "A10_Score"     "age"          "gender"
## [13] "ethnicity"     "jundice"       "austim"        "contry_of_res"
## [17] "used_app_before" "result"       "age_desc"     "relation"
## [21] "Class.ASD"

```

Check the dimensions: the dataset consists of 704 observations and 21 variables.

```
dim(df)
```

```
## [1] 704 21
```

Convert to data table

```
dt <- as.data.table(df)
```

Examine the structure of the data

```
str(dt)
```

```

## Classes 'data.table' and 'data.frame': 704 obs. of 21 variables:
## $ A1_Score      : int 1 1 1 1 1 1 0 1 1 1 ...
## $ A2_Score      : int 1 1 1 1 0 1 1 1 1 1 ...
## $ A3_Score      : int 1 0 0 0 0 1 0 1 0 1 ...
## $ A4_Score      : int 1 1 1 1 0 1 0 1 0 1 ...
## $ A5_Score      : int 0 0 1 0 0 1 0 0 1 0 ...
## $ A6_Score      : int 0 0 0 0 0 0 0 0 0 1 ...
## $ A7_Score      : int 1 0 1 1 0 1 0 0 0 1 ...
## $ A8_Score      : int 1 1 1 1 1 1 1 0 1 1 ...
## $ A9_Score      : int 0 0 1 0 0 1 0 1 1 1 ...
## $ A10_Score     : int 0 1 1 1 0 1 0 0 1 0 ...
## $ age           : chr "26" "24" "27" "35" ...
## $ gender         : chr "f" "m" "m" "f" ...
## $ ethnicity     : chr "White-European" "Latino" "Latino" "White-European" ...
## $ jundice        : chr "no" "no" "yes" "no" ...
## $ austim          : chr "no" "yes" "yes" "yes" ...
## $ contry_of_res   : chr "'United States'" "Brazil" "Spain" "'United States'" ...
## $ used_app_before: chr "no" "no" "no" "no" ...
## $ result          : int 6 5 8 6 2 9 2 5 6 8 ...
## $ age_desc        : chr "'18 and more'" "'18 and more'" "'18 and more'" "'18 and more'" ...
## $ relation         : chr "Self" "Self" "Parent" "Self" ...
## $ Class.ASD       : chr "NO" "NO" "YES" "NO" ...
## - attr(*, ".internal.selfref")=<externalptr>

```

Data preprocessing

Select variables of interest

```
dt <- dt[, c("A1_Score", "A2_Score", "A3_Score", "A4_Score", "A5_Score", "A6_Score", "A7_Score", "A8_Score")]
```

Check for missing values

```
colSums(is.na(dt))
```

```
##      A1_Score      A2_Score      A3_Score      A4_Score      A5_Score
##            0            0            0            0            0
##      A6_Score      A7_Score      A8_Score      A9_Score     A10_Score
##            0            0            0            0            0
##      gender ethnicity austim contry_of_res      result
##            0            0            0            0            0
## Class.ASD
##            0
```

Convert gender and autism to factor

```
dt[, gender := as.factor(gender)]
```

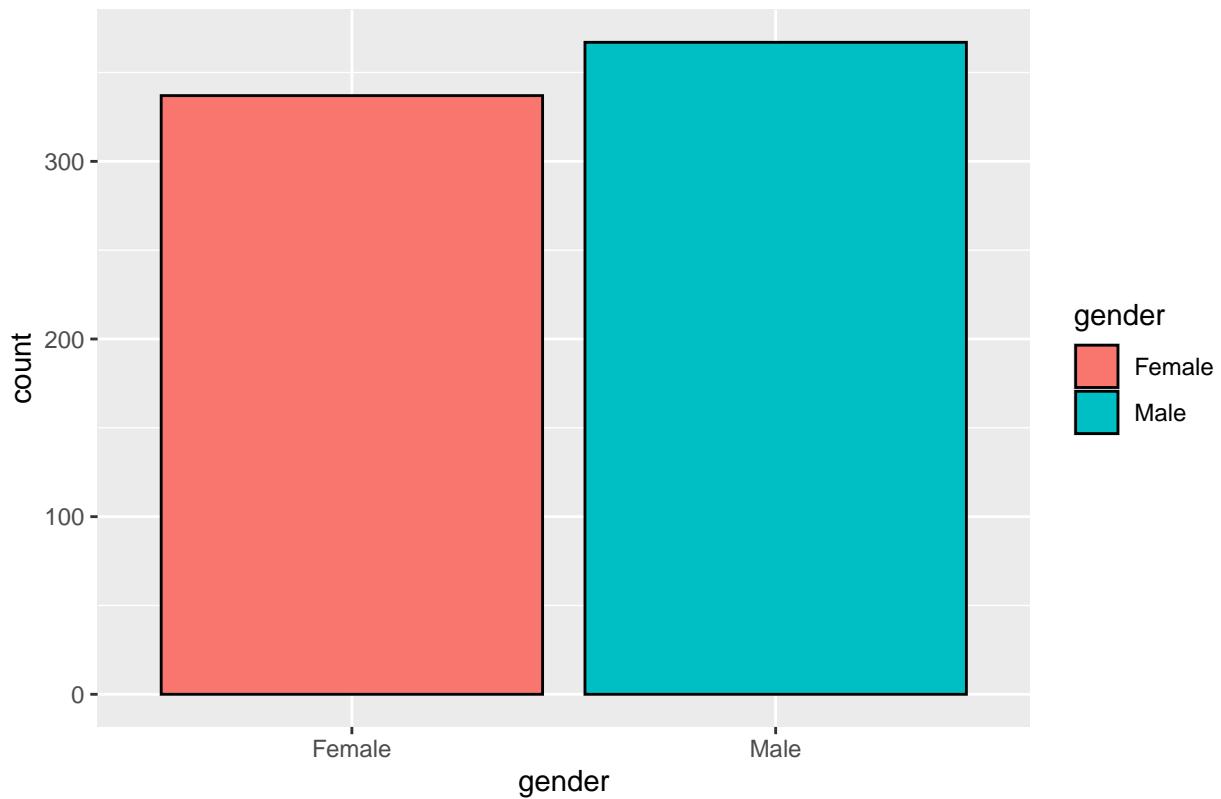
Convert class to numeric

```
dt$Class.ASD <- ifelse(dt$Class.ASD == "YES", 1, 0)
```

Distribution of gender

```
ggplot(dt, aes(x=gender, fill=gender)) +
  geom_bar(color="black") +
  scale_fill_discrete(
    labels=c("f"="Female", "m"="Male"))
  ) +
  scale_x_discrete(
    labels=c("f"="Female", "m"="Male"))
  ) +
  ggtitle("Distribution of gender among participants")
```

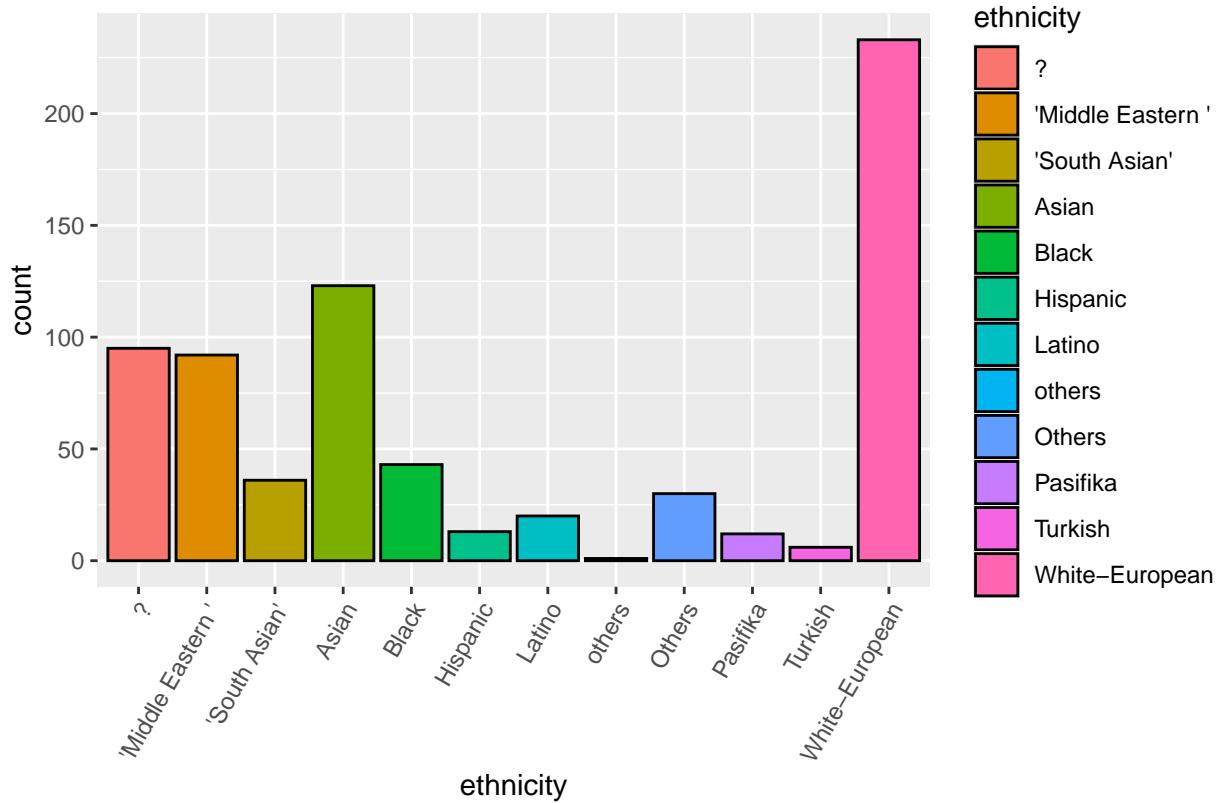
Distribution of gender among participants



Distribution of ethnicity

```
ggplot(dt, aes(x=ethnicity, fill=ethnicity)) +
  geom_bar(color="black") +
  ggtitle("Distribution of ethnicity among participants") +
  theme(axis.text.x = element_text(angle=60, hjust=1))
```

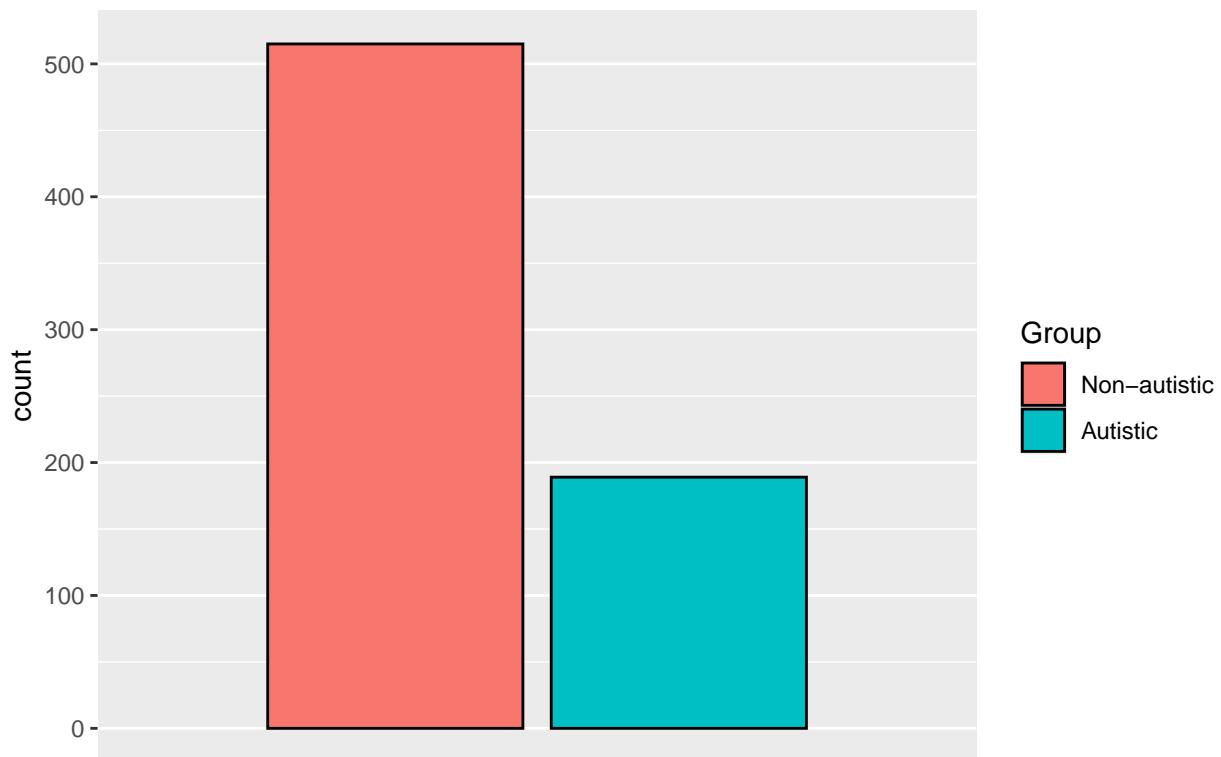
Distribution of ethnicity among participants



Distribution of autistic vs. non-autistic participants

```
ggplot(dt, aes(x=Class.ASD, fill=as.factor(Class.ASD))) +
  geom_bar(color="black") +
  labs(
    title = "Distribution of autistic vs. non-autistic participants",
    fill = "Group"
  ) +
  scale_fill_discrete(
    labels=c("0"="Non-autistic", "1"="Autistic")
  ) +
  scale_x_discrete(labels=none) +
  xlab("")
```

Distribution of autistic vs. non-autistic participants



Data Analysis

Which of the ten AQ-10 items (A1–A10) are most strongly associated with an autism diagnosis?

```
correlations <- sapply(dt[, 1:10], function(x) cor(x, dt[[16]]))

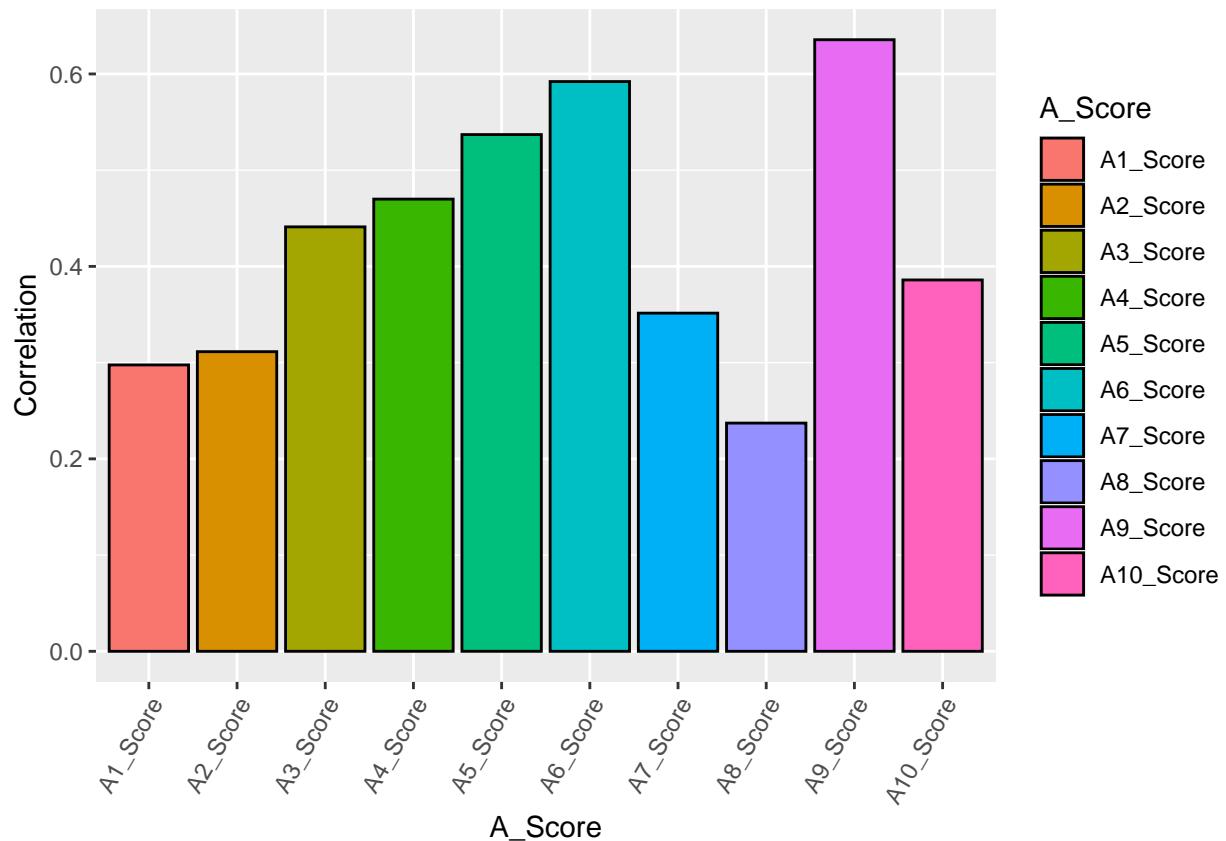
correlations

##  A1_Score  A2_Score  A3_Score  A4_Score  A5_Score  A6_Score  A7_Score  A8_Score
## 0.2976276 0.3113817 0.4410737 0.4699452 0.5370043 0.5920910 0.3514286 0.2371606
##  A9_Score A10_Score
## 0.6355758 0.3859171

corr <- data.frame(
  A_Score=names(correlations),
  Correlation=as.numeric(correlations)
)

corr$A_Score <- factor(corr$A_Score, levels = corr$A_Score)
```

```
ggplot(corr, aes(x=A_Score, y=Correlation, fill=A_Score)) +
  geom_bar(stat = "identity", color="black") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



Is there a difference in the distribution of autism diagnoses between genders?

Perform ANOVA test to examine the difference between genders

```
anova_result <- aov(as.numeric(Class.ASD) ~ as.factor(gender), dt)
```

```
summary(anova_result)
```

```
##                               Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(gender)      1   0.89  0.8932   4.565  0.033 *
## Residuals              702 137.37  0.1957
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

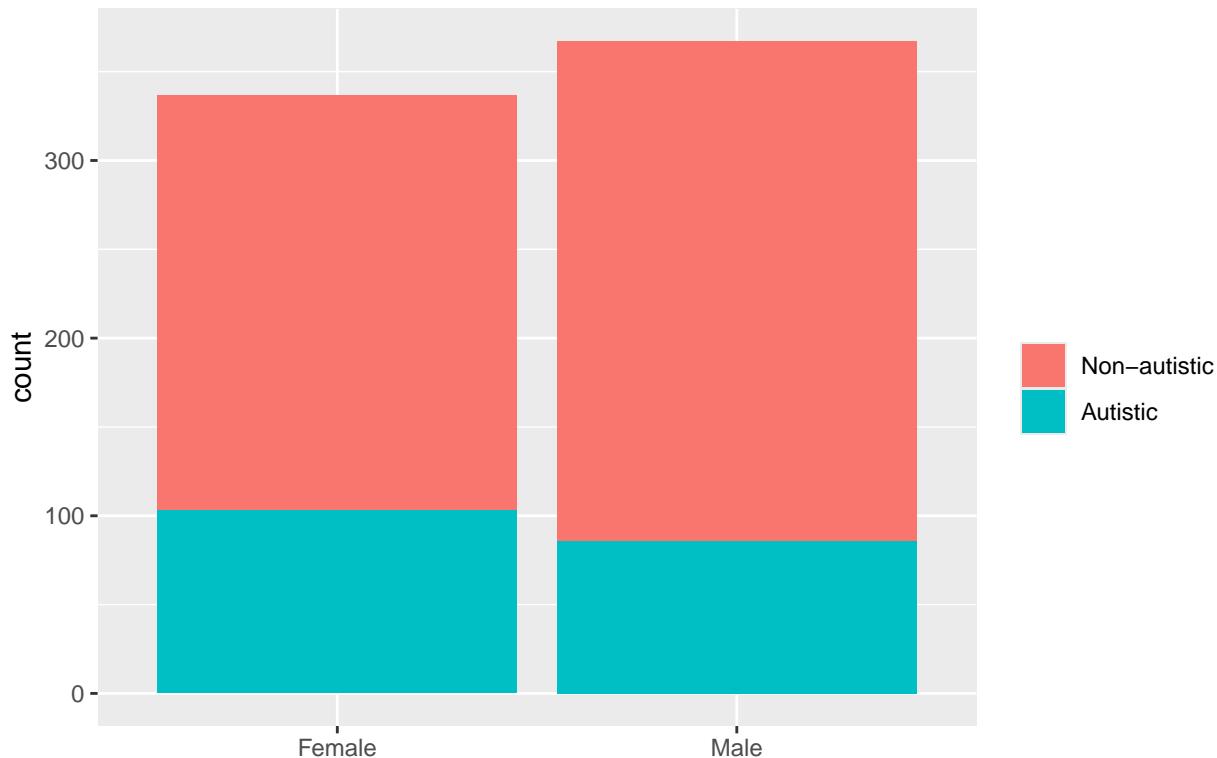
The ANOVA results indicate a statistically significant difference in the number of autism diagnoses between genders (p-value = 0.033).

```

ggplot(dt, aes(
  x = factor(gender, levels=c("f", "m"), labels=c("Female", "Male")),
  fill = factor(Class.ASD, levels=c(0, 1), labels=c("Non-autistic", "Autistic")))) +
  geom_bar(position = "stack") +
  labs(
    x = "",
    fill = "",
    title = "Distribution of autism diagnoses between genders"
)

```

Distribution of autism diagnoses between genders



The graph indicates that among participants women were statistically more likely to be diagnosed with autism compared to men.

Do males and females differ in their average AQ-10 total scores?

Perform ANOVA test:

```
anova_result <- aov(as.numeric(result) ~ as.factor(gender), dt)
```

```
summary(anova_result)
```

```

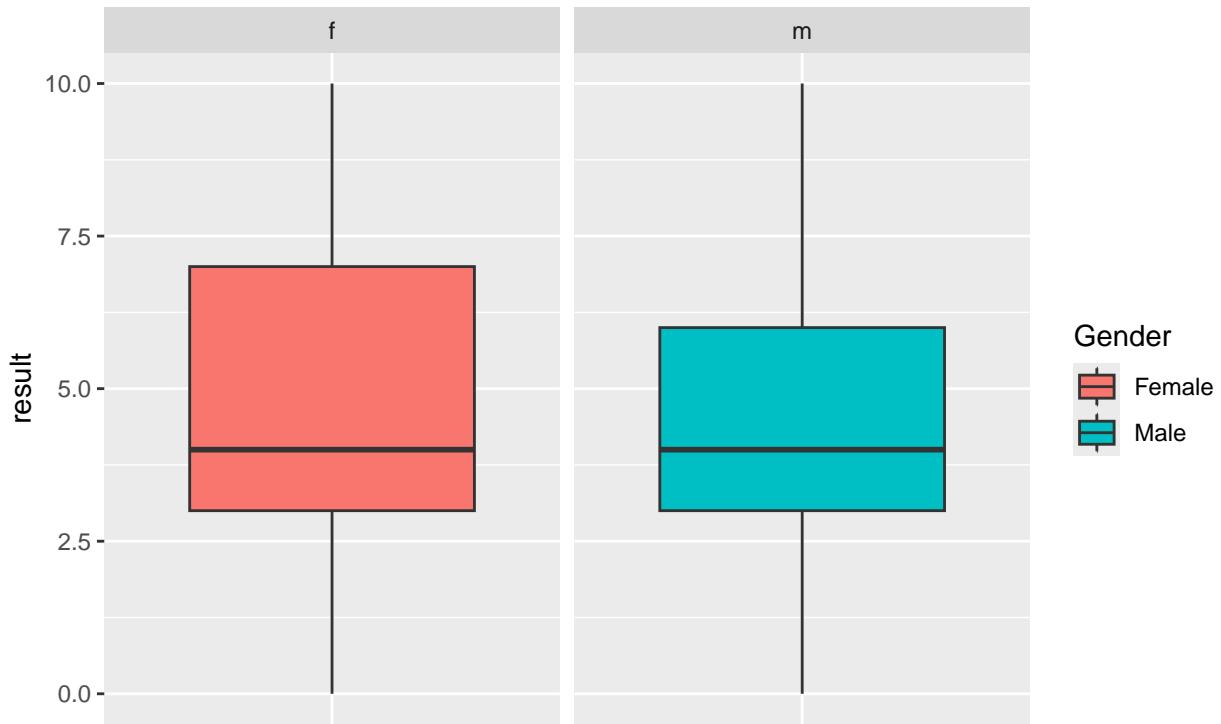
##                               Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(gender)      1     8   7.845   1.254  0.263
## Residuals              702  4391   6.255

```

There is no statistically significant difference in the distribution of results between genders ($p = 0.263$).

```
ggplot(dt, aes(x="", y=result, fill=factor(gender, levels=c("f", "m"), labels=c("Female", "Male")))) +  
  geom_boxplot() +  
  facet_wrap(~gender) +  
  labs(  
    fill="Gender",  
    x="",  
    title="AQ-10 total scores distribution among all participants"  
)
```

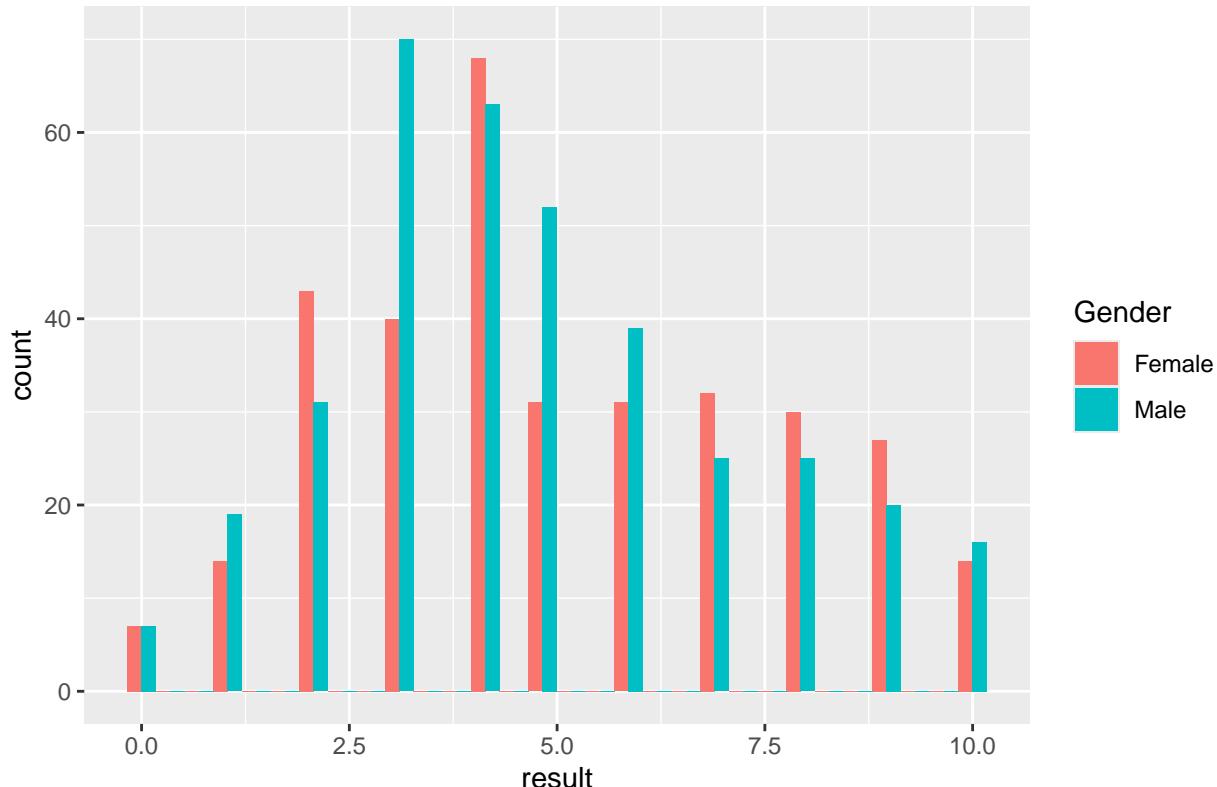
AQ-10 total scores distribution among all participants



```
ggplot(dt, aes(x=result, fill=factor(gender, levels=c("f", "m"), labels=c("Female", "Male")))) +  
  geom_histogram(position = "dodge") +  
  labs(  
    fill="Gender",  
    title="AQ-10 total scores distribution among all participants"  
)
```

'stat_bin()' using 'bins = 30'. Pick better value 'binwidth'.

AQ-10 total scores distribution among all participants



The plots illustrate that although the third quartile (Q3) for women is slightly higher - indicating that women tend to achieve marginally higher scores - the median scores are comparable across genders.

Does the prevalence of autism vary across different ethnic groups?

```
unique(dt$ethnicity)
```

```
## [1] "White-European"      "Latino"           "?"  
## [4] "Others"             "Black"            "Asian"  
## [7] "'Middle Eastern'"   "'Pasifika'"       "'South Asian'"  
## [10] "Hispanic"          "Turkish"         "others"
```

Remove groups "Others", "others" and "?"

```
filteredDT <- dt[ethnicity != "Others" & ethnicity != "?" & ethnicity != "others"]
```

```
dim(filteredDT)
```

```
## [1] 578 16
```

Perform ANOVA test:

```

anova_result <- aov(as.numeric(result) ~ as.factor(ethnicity), filteredDT)

summary(anova_result)

##                                Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(ethnicity)     8   526   65.79   11.69 1.88e-15 ***
## Residuals                  569   3203    5.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The results of the ANOVA test indicate statistical significance in the distribution of results across ethnic groups (marked as ***).

```

filteredDT[, .(
  meanResult = mean(result),
  minResult = min(result),
  maxResult = max(result),
  medianResult = median(result),
  numberParticipants = .N
), by=ethnicity][order(-meanResult)]

##           ethnicity meanResult minResult maxResult medianResult
## 1:        Latino    6.350000      3       10       6.5
## 2: White-European    6.034335      0       10       6.0
## 3:        Black     5.325581      0       10       6.0
## 4:    Hispanic     5.076923      1       9       6.0
## 5:    Turkish      4.333333      2       9       4.0
## 6:       Asian      4.276423      0       10       4.0
## 7: 'Middle Eastern'    3.978261      0       10       4.0
## 8: 'South Asian'     3.777778      0       8       4.0
## 9:     Pasifika     3.666667      1       10       3.5
##   numberParticipants
## 1:                      20
## 2:                     233
## 3:                      43
## 4:                      13
## 5:                      6
## 6:                     123
## 7:                      92
## 8:                      36
## 9:                      12

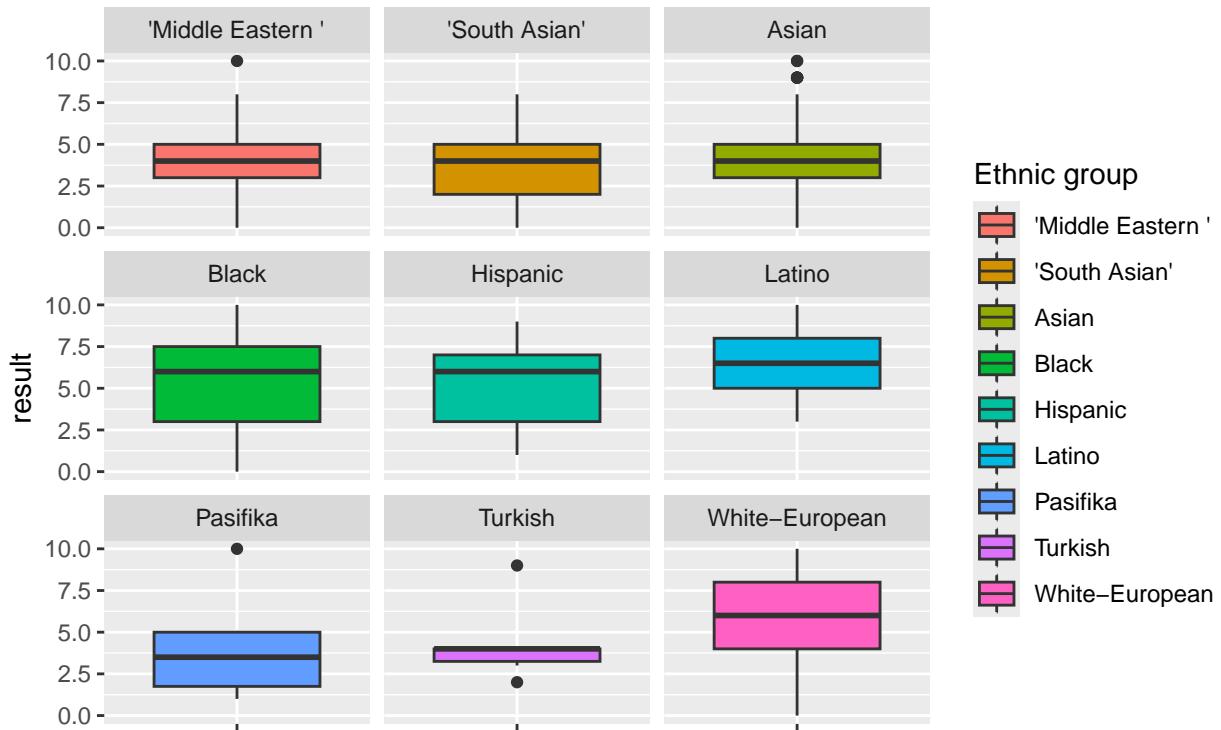
```

```

ggplot(filteredDT, aes(x="", y=result, fill=factor(ethnicity))) +
  geom_boxplot() +
  facet_wrap(~ethnicity) +
  labs(
    fill="Ethnic group",
    x="",
    title="AQ-10 total scores distribution among all participants"
  )

```

AQ-10 total scores distribution among all participants



Further analysis indicated that participants identifying as Latino, White-European, and Black achieved the highest scores on the AQ10 screening, whereas those identifying as Pacifica, South Asian, and Middle Eastern obtained the lowest scores, suggesting they are less likely to score high on the autism screening test.

Is having a family history of autism associated with higher AQ-10 scores?

```
dt []
```

```
##      A1_Score A2_Score A3_Score A4_Score A5_Score A6_Score A7_Score A8_Score
##      <int>    <int>    <int>    <int>    <int>    <int>    <int>    <int>
## 1:       1       1       1       1       0       0       1       1
## 2:       1       1       0       1       0       0       0       1
## 3:       1       1       0       1       1       0       1       1
## 4:       1       1       0       1       0       0       1       1
## 5:       1       0       0       0       0       0       0       1
## ---
## 700:      0       1       0       1       1       0       1       1
## 701:      1       0       0       0       0       0       0       1
## 702:      1       0       1       1       1       0       1       1
## 703:      1       0       0       1       1       0       1       0
## 704:      1       0       1       1       1       0       1       1
##      A9_Score A10_Score gender      ethnicity austim   contry_of_res result
##      <int>    <int> <fctr>      <char> <char>      <char>    <int>
## 1:       0       0     f White-European     no 'United States'     6
## 2:       0       1     m Latino           yes Brazil        5
```

```

##   3:      1      1     m      Latino    yes       Spain     8
##   4:      0      1     f White-European  yes 'United States' 6
##   5:      0      0     f                  ?    no        Egypt   2
##   ---
## 700:      1      1     f White-European  no        Russia    7
## 701:      0      1     m      Hispanic  no        Mexico    3
## 702:      0      1     f                  ?    no        Russia    7
## 703:      1      1     m  'South Asian'  no        Pakistan  6
## 704:      1      1     f White-European  no        Cyprus    8
##   Class.ASD
##   <num>
##   1:      0
##   2:      0
##   3:      1
##   4:      0
##   5:      0
##   ---
## 700:      1
## 701:      0
## 702:      1
## 703:      0
## 704:      1

```

Perform ANOVA test:

```

dt$austim <- as.factor(dt$austim)

unique(dt$austim)

## [1] no  yes
## Levels: no yes

anova_result <- aov(as.numeric(result) ~ as.factor(austim), data = dt)

summary(anova_result)

##                               Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(austim)      1   162   162.22  26.88 2.84e-07 ***
## Residuals             702  4237    6.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The results of the ANOVA test indicate statistical significance of the family history of autism (marked as ***).

```

dt[, .(
  meanResult = mean(result),
  minResult = min(result),
  maxResult = max(result),
  medianResult = median(result),
  numberParticipants = .N
), by=austim][order(-meanResult)]

```

```

##      austim meanResult minResult maxResult medianResult number0fParticipants
##      <fctr>     <num>     <int>     <int>     <num>             <int>
## 1: yes    6.120879        1       10        6            91
## 2: no     4.690049        0       10        4           613

```

```

ggplot(filteredDT, aes(x="", y=result, fill=factor(austim))) +
  geom_boxplot() +
  facet_wrap(~austim) +
  labs(
    fill="Family history",
    x="",
    title="Scores distribution among participants with and without family history"
)

```



Summary

The analysis explored factors associated with autism and AQ-10 screening results. Among the individual AQ-10 items, A6 (“I know how to tell if someone listening to me is getting bored”) and A9 (“I find it easy to work out what someone is thinking or feeling just by looking at their face”) showed the strongest correlation with autism diagnosis. In contrast, A8 (“I like to collect information about categories or things”) was the least significant question in relation to autism scores.

No statistically significant difference was found between genders in the distribution of autism diagnoses or overall AQ-10 scores, although women were slightly more represented among those diagnosed.

Ethnicity showed a significant effect: participants identifying as Latino, White-European, or Black tended to achieve higher AQ-10 scores, while those identifying as Middle Eastern, South Asian, or Pacifica had lower scores, indicating a lower likelihood of scoring high on the autism screening.

Finally, a strong association was observed between family history of autism and higher AQ-10 results, with ANOVA confirming this factor as highly significant.