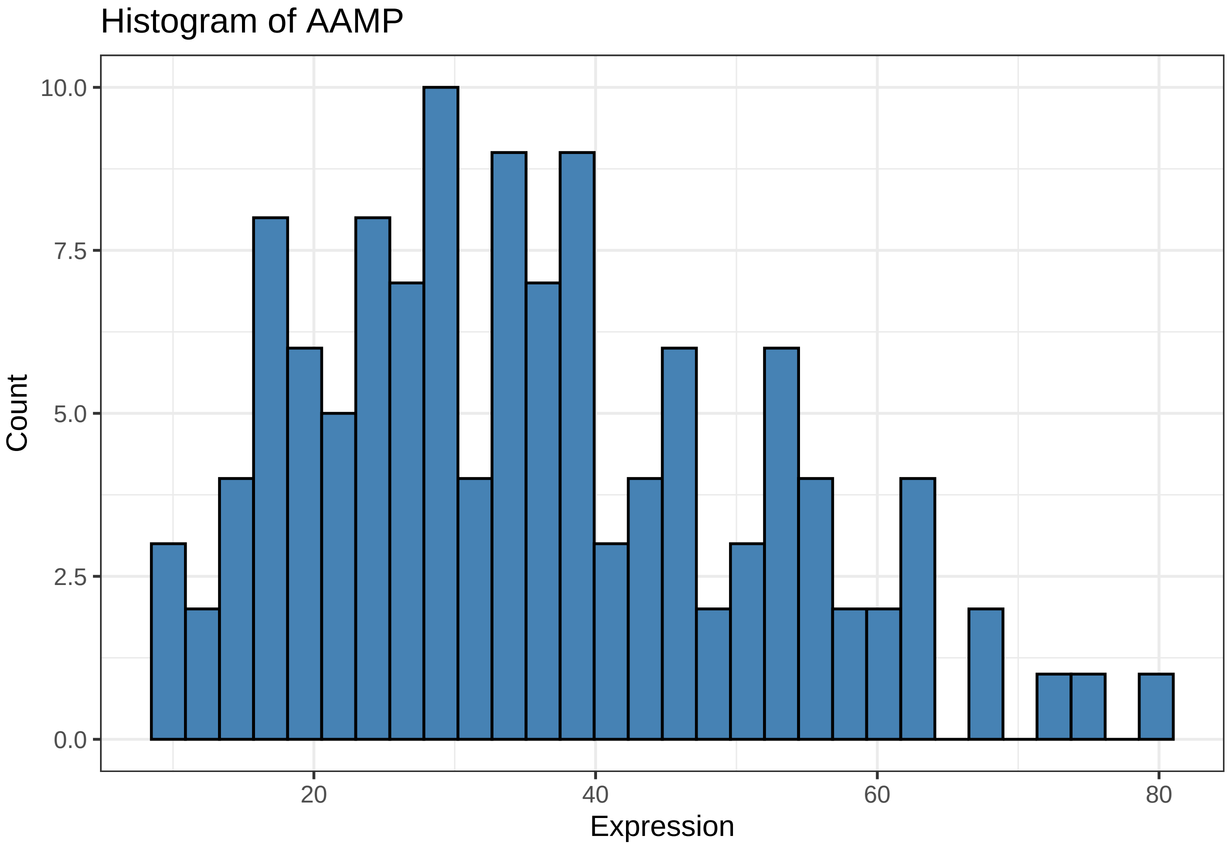
```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)

# read expression and metadata
genes <- read.csv("~/Fundations of Data Science 103/Submission 1/QBS103_GSE157103_genes.csv", check.name
meta <- read.csv("~/Fundations of Data Science 103/Submission 1/QBS103_GSE157103_series_matrix-1.csv")

# rename gene column
colnames(genes)[1] <- "gene"

# reshape gene table
gene_long <- genes %>%
  pivot_longer(-gene, names_to = "participant_id", values_to = "expression") %>%
  pivot_wider(names_from = gene, values_from = expression)

# merge with metadata
merged <- inner_join(meta, gene_long, by = "participant_id")
merged$age <- as.numeric(merged$age)
```

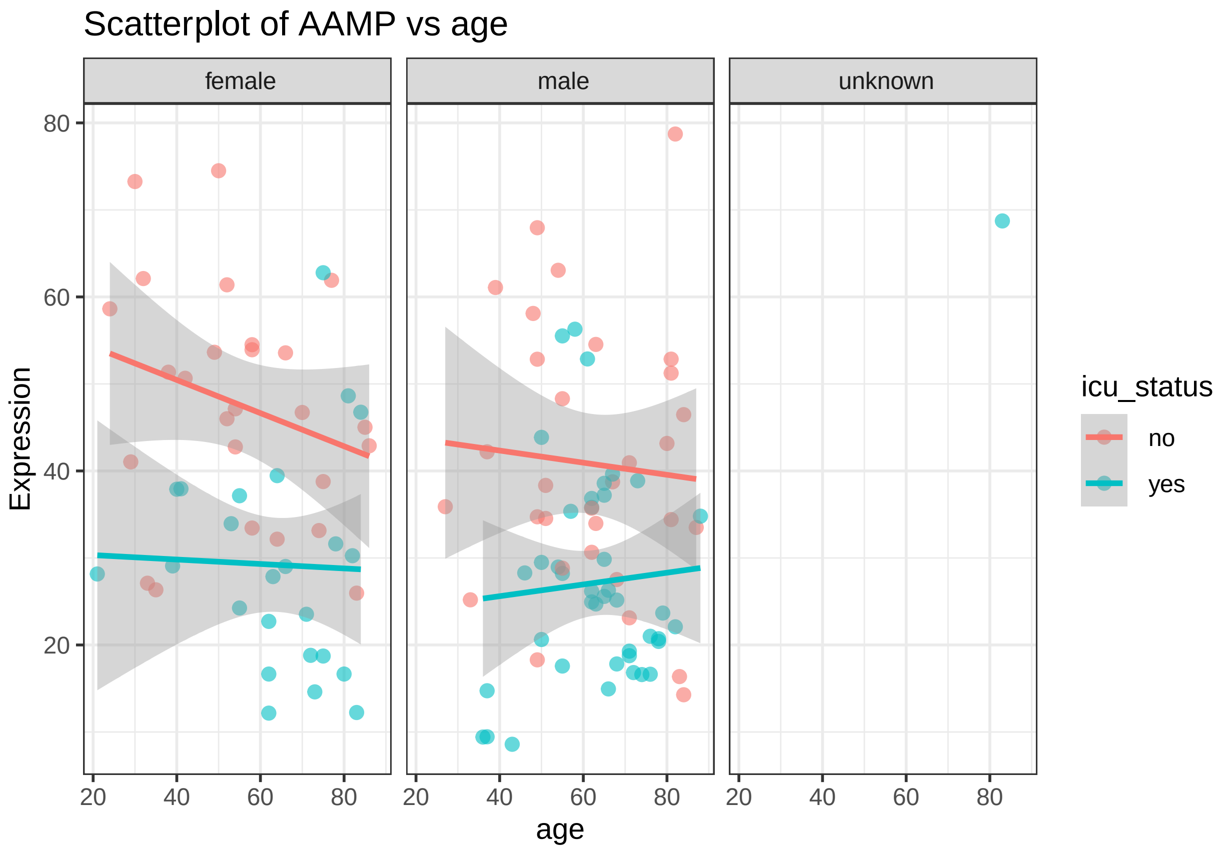
```
## Warning: NAs introduced by coercion
```

```r
# clean missing or unknown values
merged <- merged %>%
  filter(!is.na(age),
         !is.na(sex), sex != "unknown",
         !is.na(icu_status), icu_status != "unknown")

# define function to return three plots for one gene
plot_all_types <- function(df, gene, cont_cov, cat_cov1, cat_cov2) {
  p1 <- ggplot(df, aes(x = .data[[gene]])) +
    geom_histogram(bins = 30, fill = "steelblue", color = "black") +
    labs(title = paste("Histogram of", gene), x = "Expression", y = "Count") +
    theme_bw()

  p2 <- ggplot(df, aes(x = .data[[cont_cov]], y = .data[[gene]], color = .data[[cat_cov2]])) +
    geom_point(alpha = 0.6, size = 2) +
    geom_smooth(method = "lm", se = TRUE) +
    facet_wrap(~.data[[cat_cov1]]) +
    labs(title = paste("Scatterplot of", gene, "vs", cont_cov),
         x = cont_cov, y = "Expression", color = cat_cov2) +
    theme_bw()

  p3 <- ggplot(df, aes(x = .data[[cat_cov1]], y = .data[[gene]], fill = .data[[cat_cov2]])) +
    geom_boxplot() +
    labs(title = paste("Boxplot of", gene),
         x = cat_cov1, y = "Expression", fill = cat_cov2) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  return(list(histogram = p1, scatter = p2, boxplot = p3))
}

# test one gene for question 1
test_plots <- plot_all_types(
  df = merged,
  gene = "AAMP",
  cont_cov = "age",
  cat_cov1 = "sex",
  cat_cov2 = "icu_status"
)

print(test_plots$histogram)
```

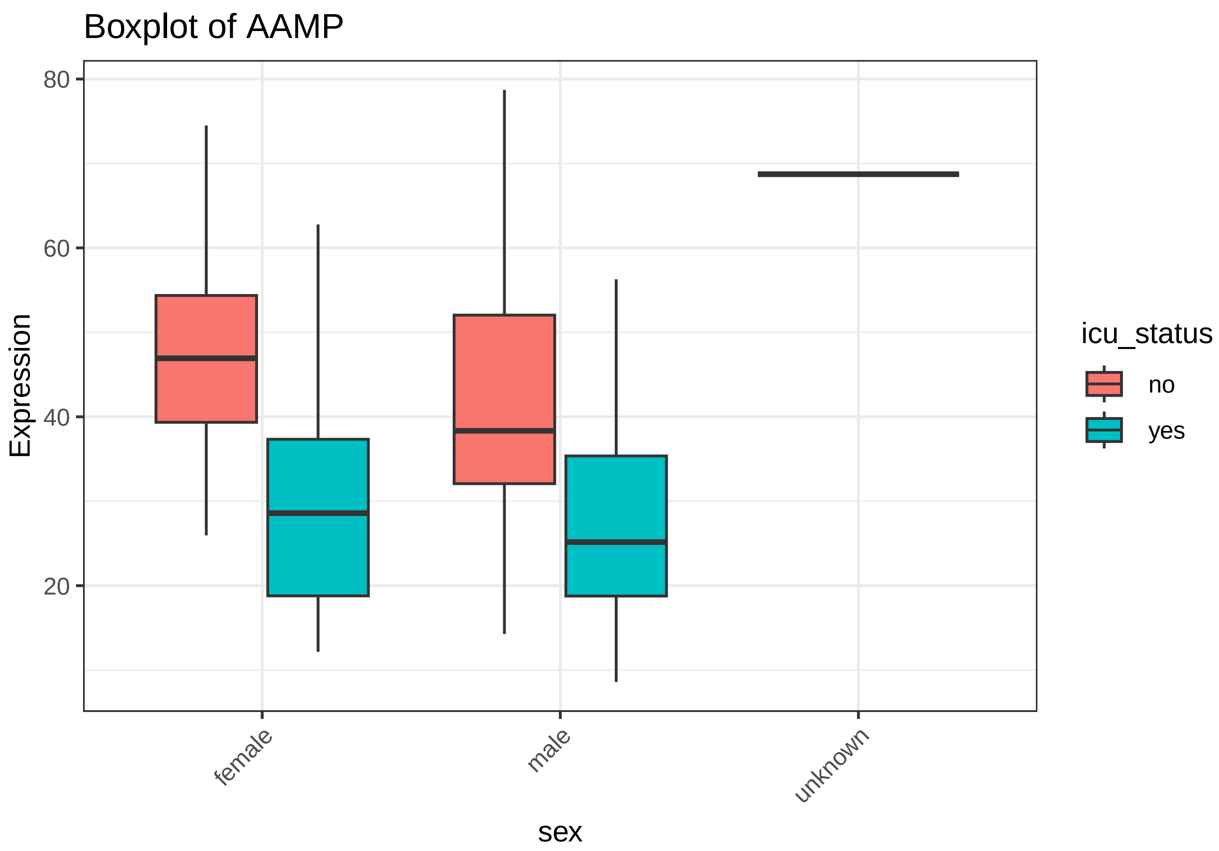


```
print(test_plots$scatter)
```

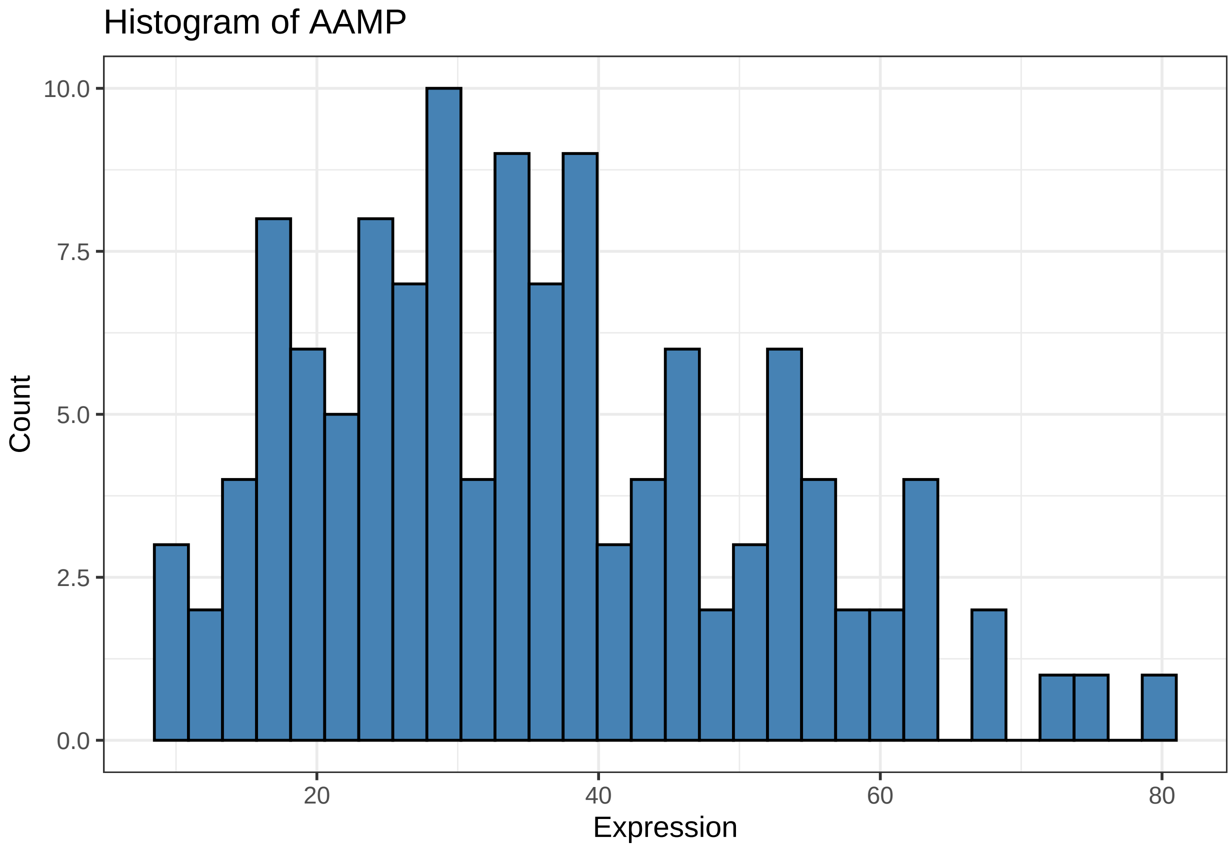
```
## 'geom_smooth()' using formula = 'y ~ x'
```



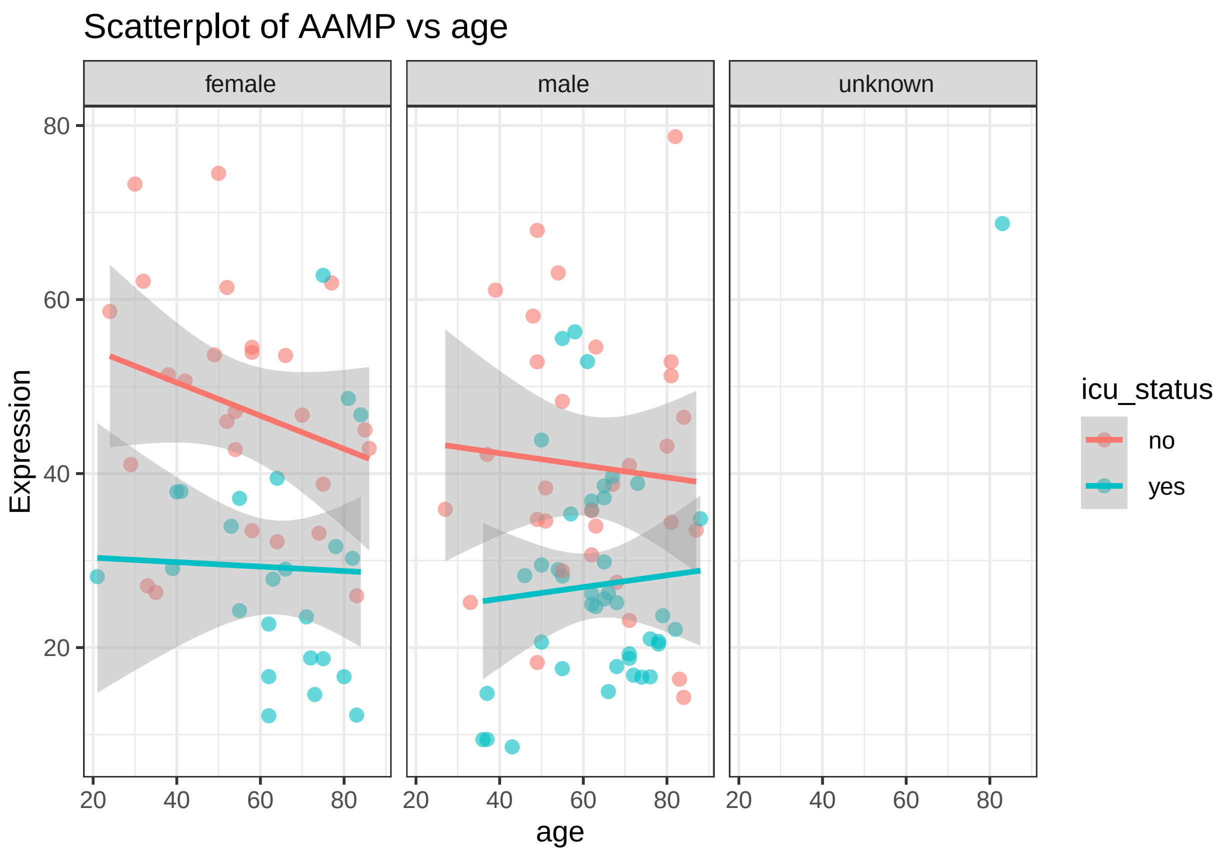
```
print(test_plots$boxplot)
```

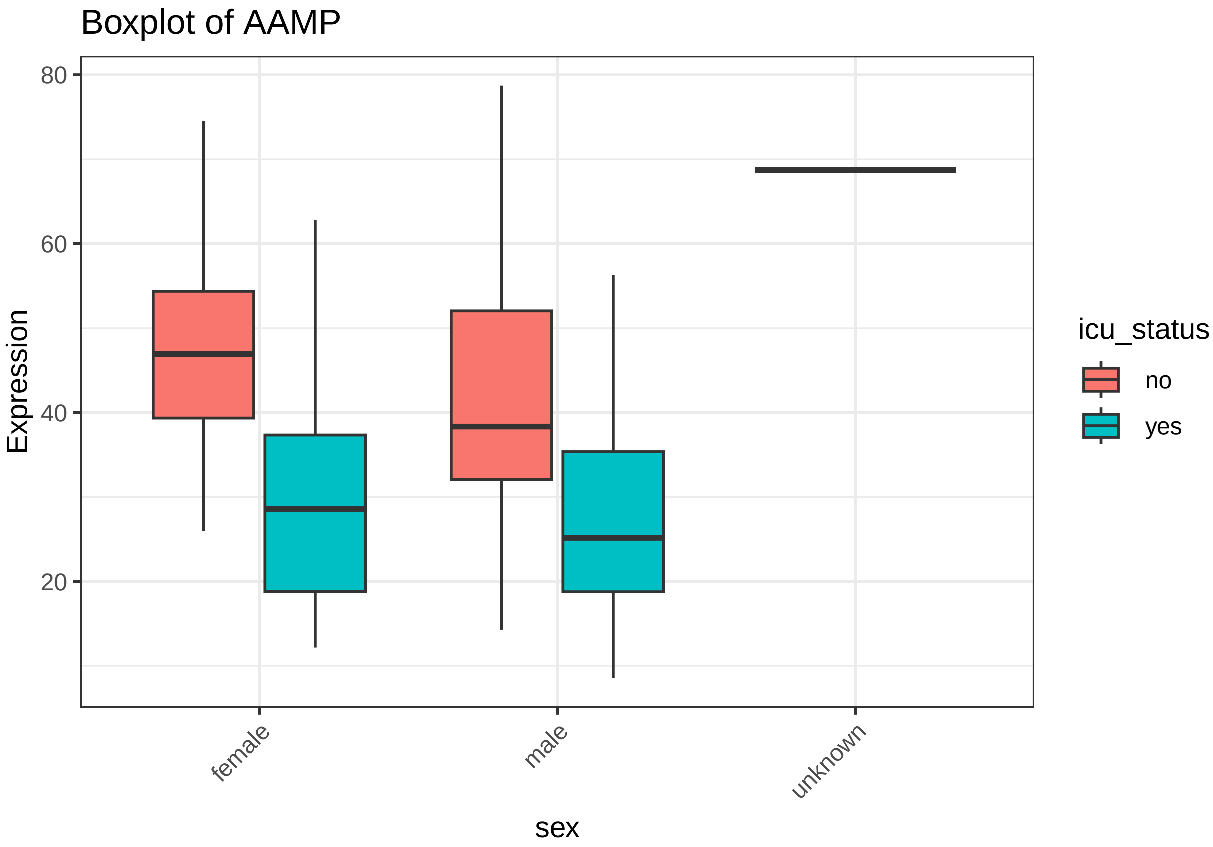


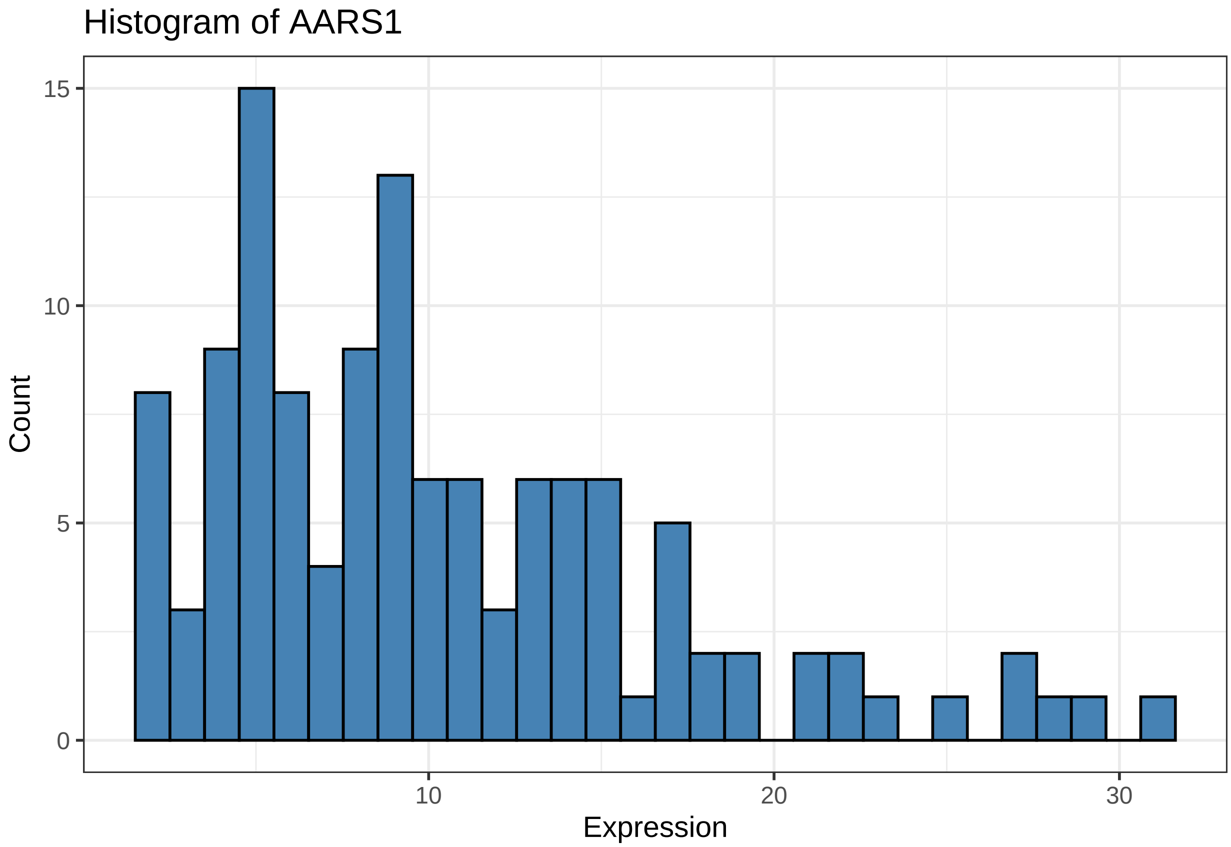
```r
# loop for question 2
selected_genes <- c("AAMP", "AARS1", "AARS2")
for (gene in selected_genes) {
  plots <- plot_all_types(
    df = merged,
    gene = gene,
    cont_cov = "age",
    cat_cov1 = "sex",
    cat_cov2 = "icu_status"
  )
  print(plots$histogram)
  print(plots$scatter)
  print(plots$boxplot)
}
```



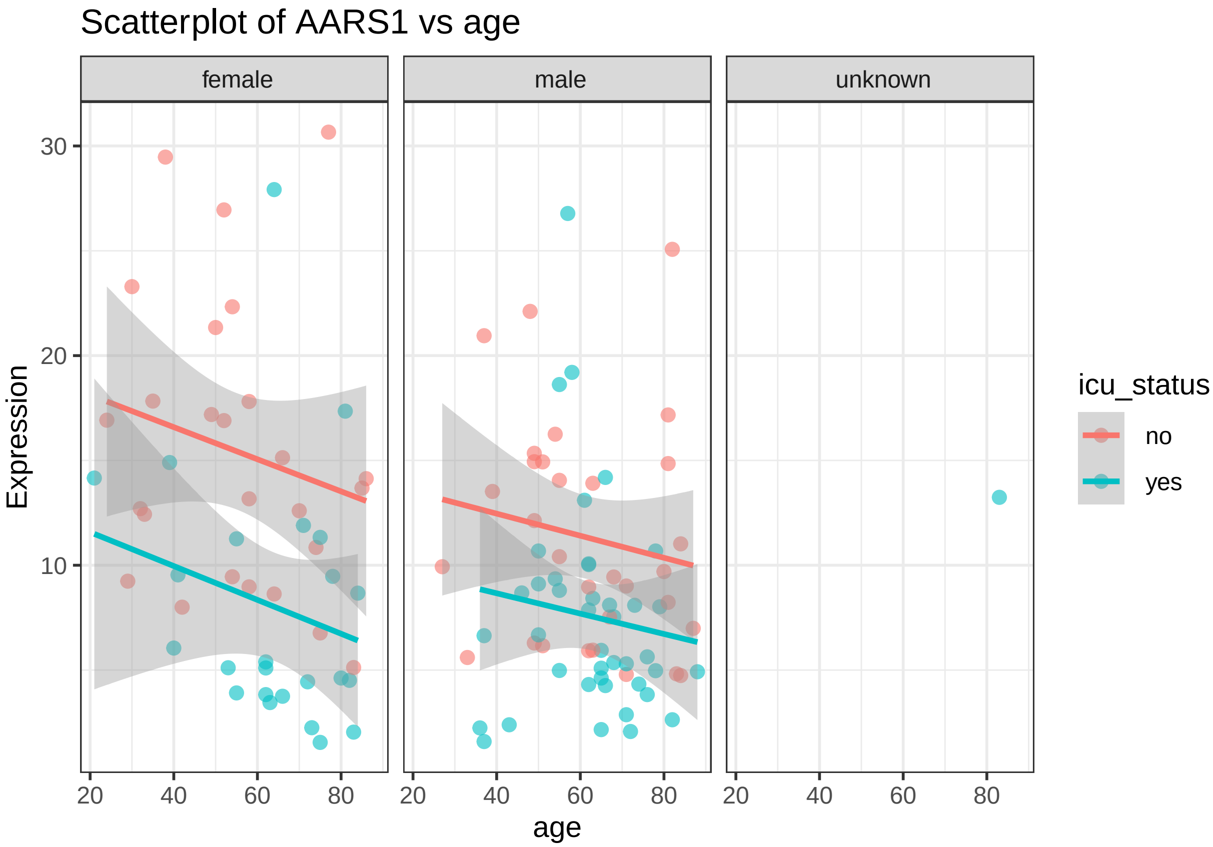
```
## `geom_smooth()` using formula = 'y ~ x'
```

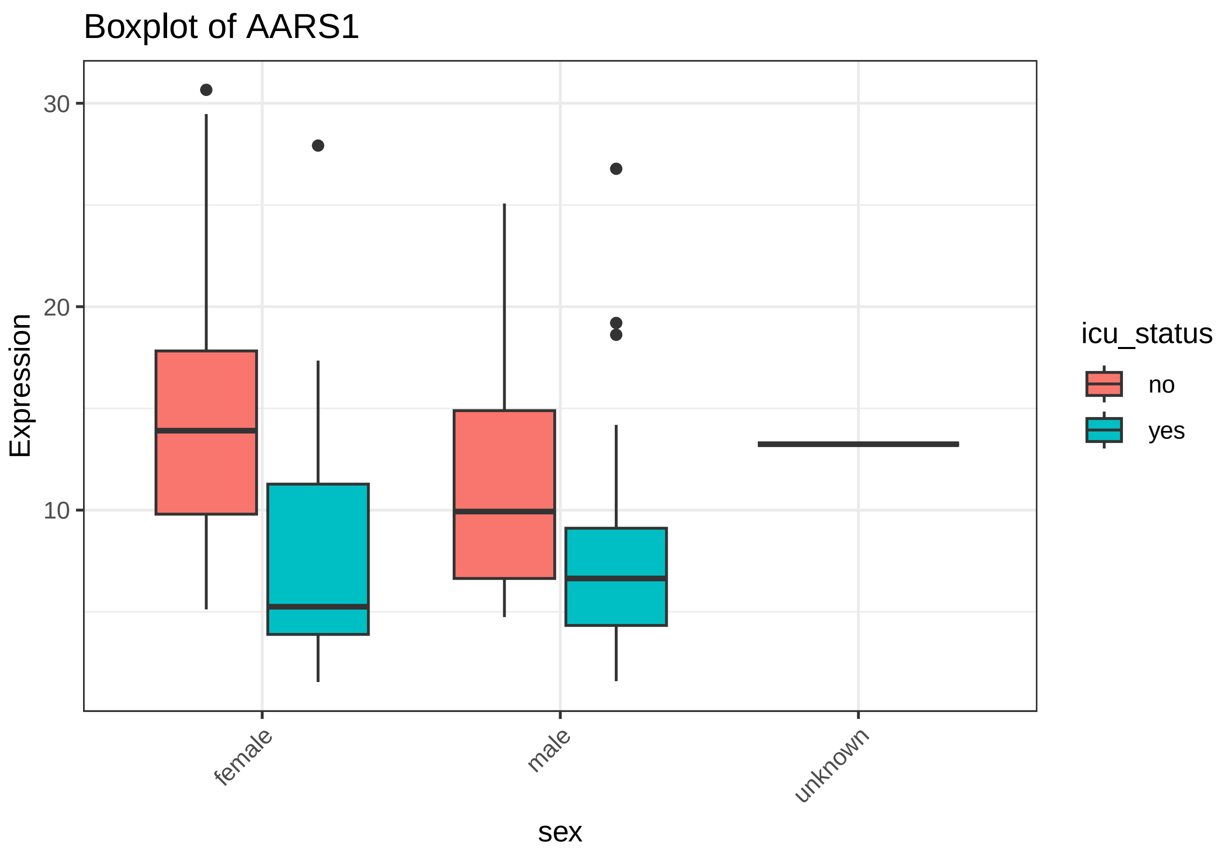

Scatterplot of AAMP vs age
female
male
unknown
Expression
80
60
40
20
age
20
40
60
80
icu_status
no
yes

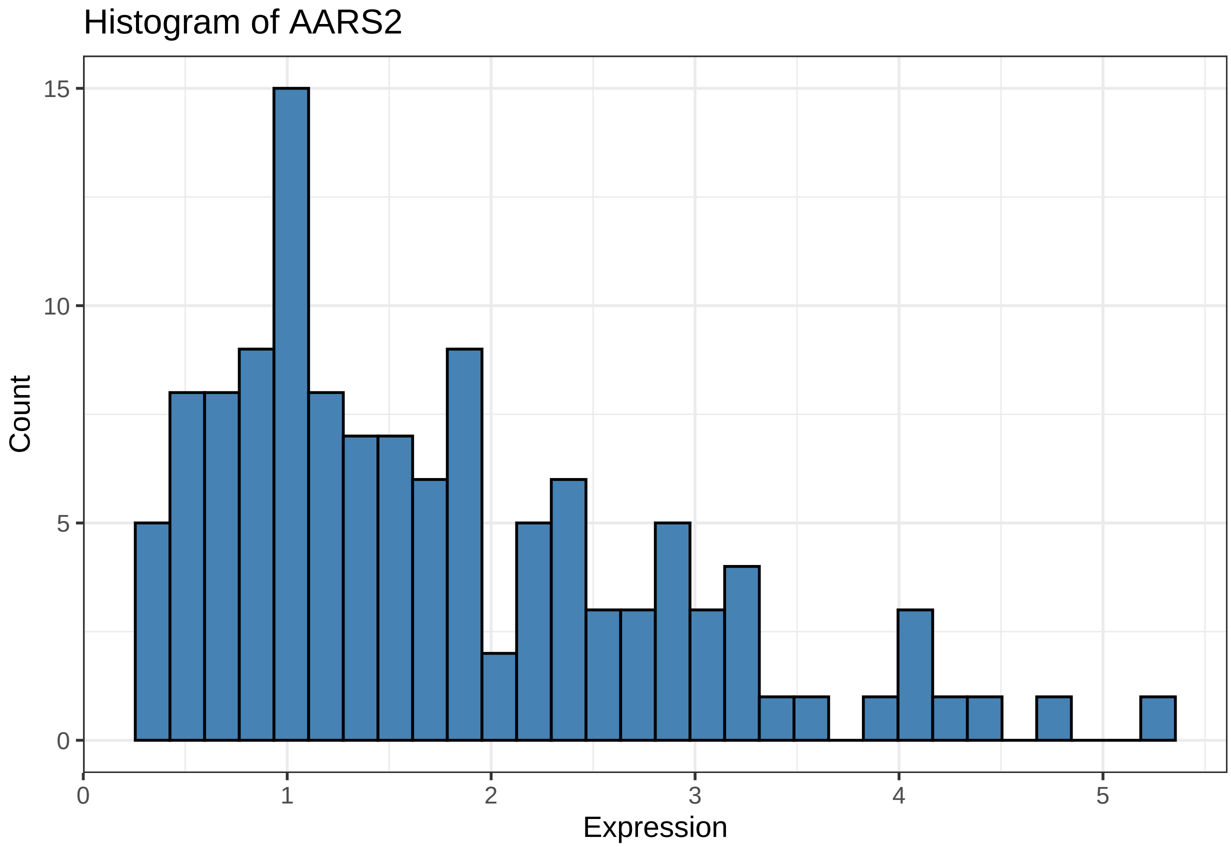

Boxplot of AAMP
80
60
40
20
Expression
female
male
unknown
sex
icu_status
no
yes



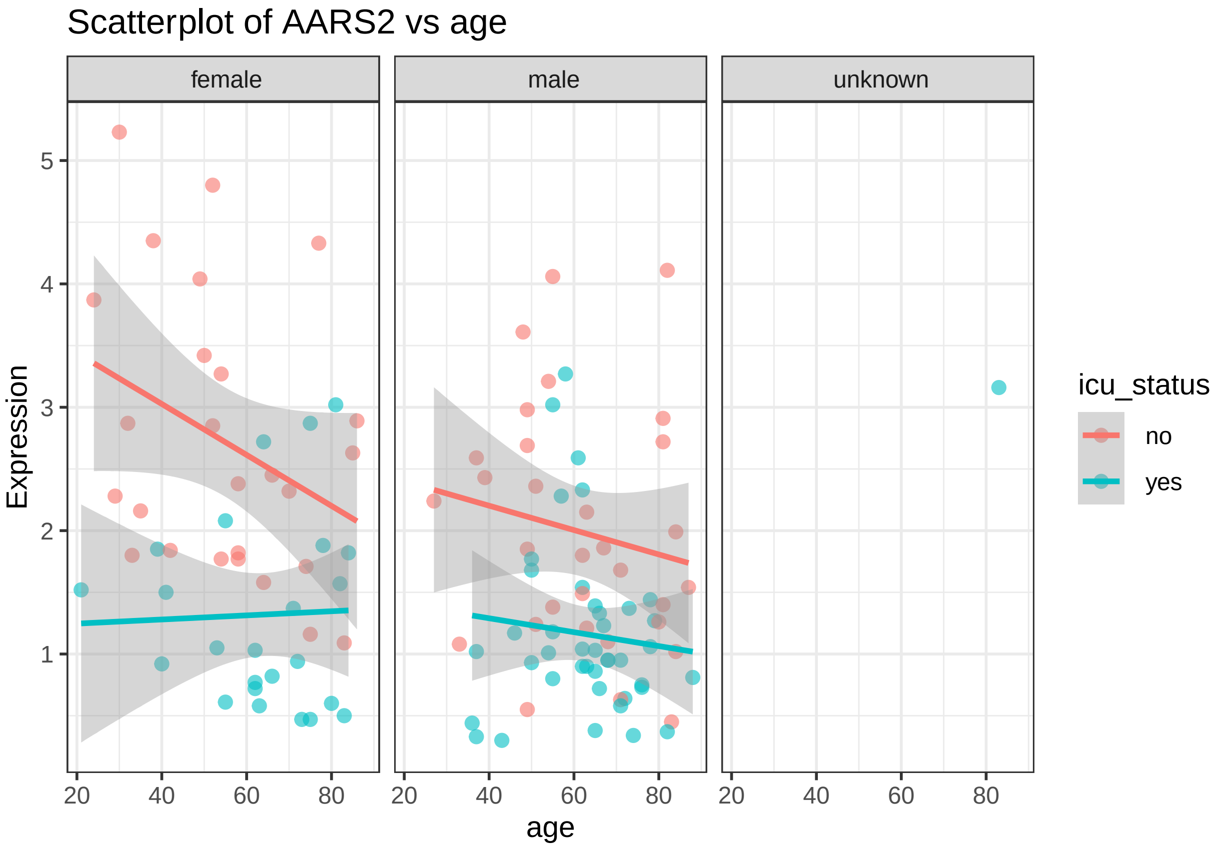
```
## `geom_smooth()` using formula = 'y ~ x'
```

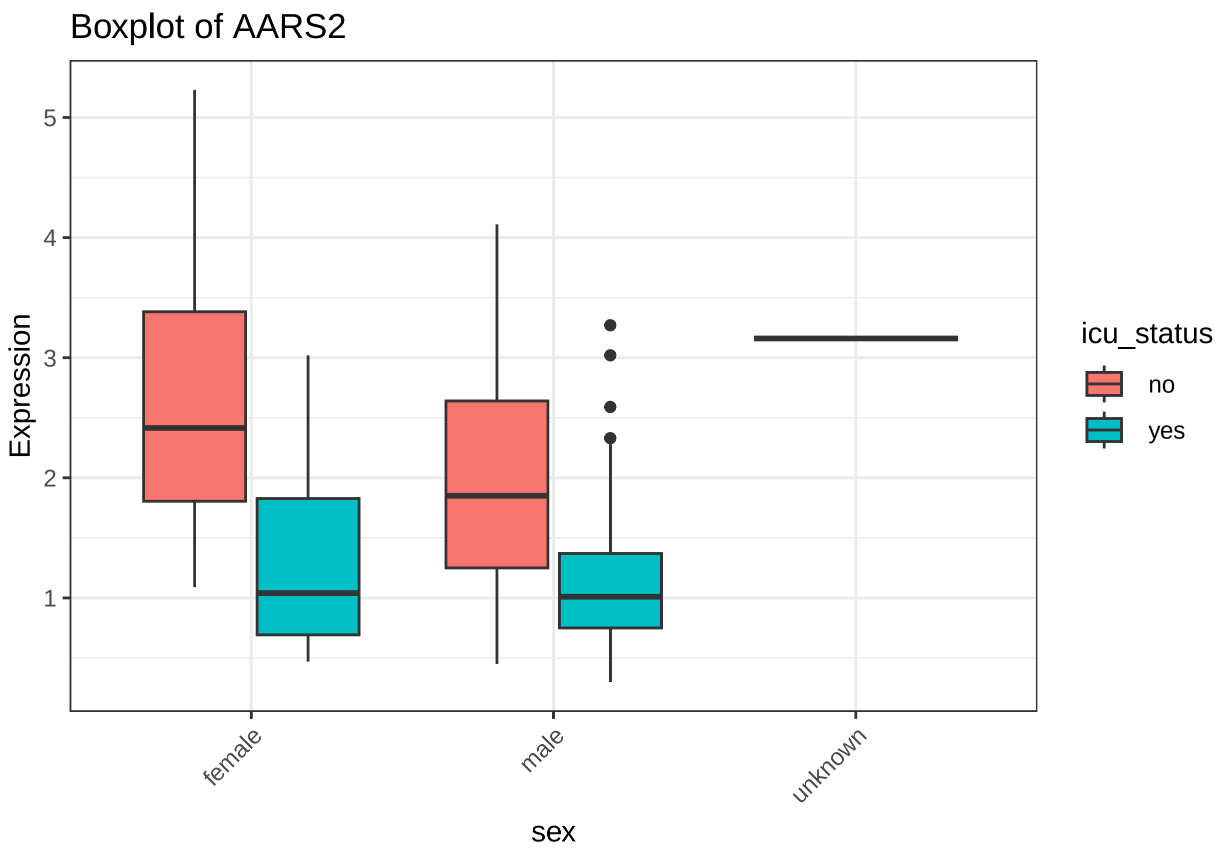

Scatterplot of AARS1 vs age
female
male
unknown
Expression
30
20
10
20
40
60
80
age
icu_status
no
yes


Boxplot of AARS1
Expression
30
20
10
female
male
unknown
sex
icu_status
no
yes



```
## `geom_smooth()` using formula = 'y ~ x'
```


Scatterplot of AARS2 vs age
female
male
unknown
Expression
5
4
3
2
1
20
40
60
80
age
icu_status
no
yes


Boxplot of AARS2
Expression
5
4
3
2
1
female
male
unknown
sex
icu_status
no
yes