Submission 2

Angelique Cortez

2024-08-08

```
set.seed(1001001)
```

```
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
```

Build a function to create the plots you made for Presentation 1, incorporating any feedback you received on your submission. Your functions should take the following input: (1) the name of the data frame, (2) a list of 1 or more gene names, (3) 1 continuous covariate, and (4) two categorical covariates (10 pts) Select 2 additional genes (for a total of 3 genes) to look at and implement a loop to generate your figures using the function you created (10 pts)

```
combined_data <- read.csv("~/Dropbox (Dartmouth College)/Dartmouth Classes/Summer2024/Intro to data ana
```

#I had issues removing the unknown age with the function, my goal for the next submission is to figure out how by next week

```
# Write a function that can recreate the plots from submission one
generate_plots <- function(data, genes, continuous_covariate, categorical_covariate1, categorical_covari

  for (gene in genes) {
    # Preprocess the data for the gene of interest
    gene_data <- data %>%
      filter(gene == !!gene) %>%
      mutate(
        # Extract age using a regular expression and convert to numeric
        !!continuous_covariate := as.numeric(sub(".*_(\\d{2})y_.*", "\\1", participant_id)),
        # Extract sex using a regular expression (assuming categorical_covariate1 is sex)
        !!categorical_covariate1 := sub(".*_(male|female)_.*", "\\1", participant_id)
```

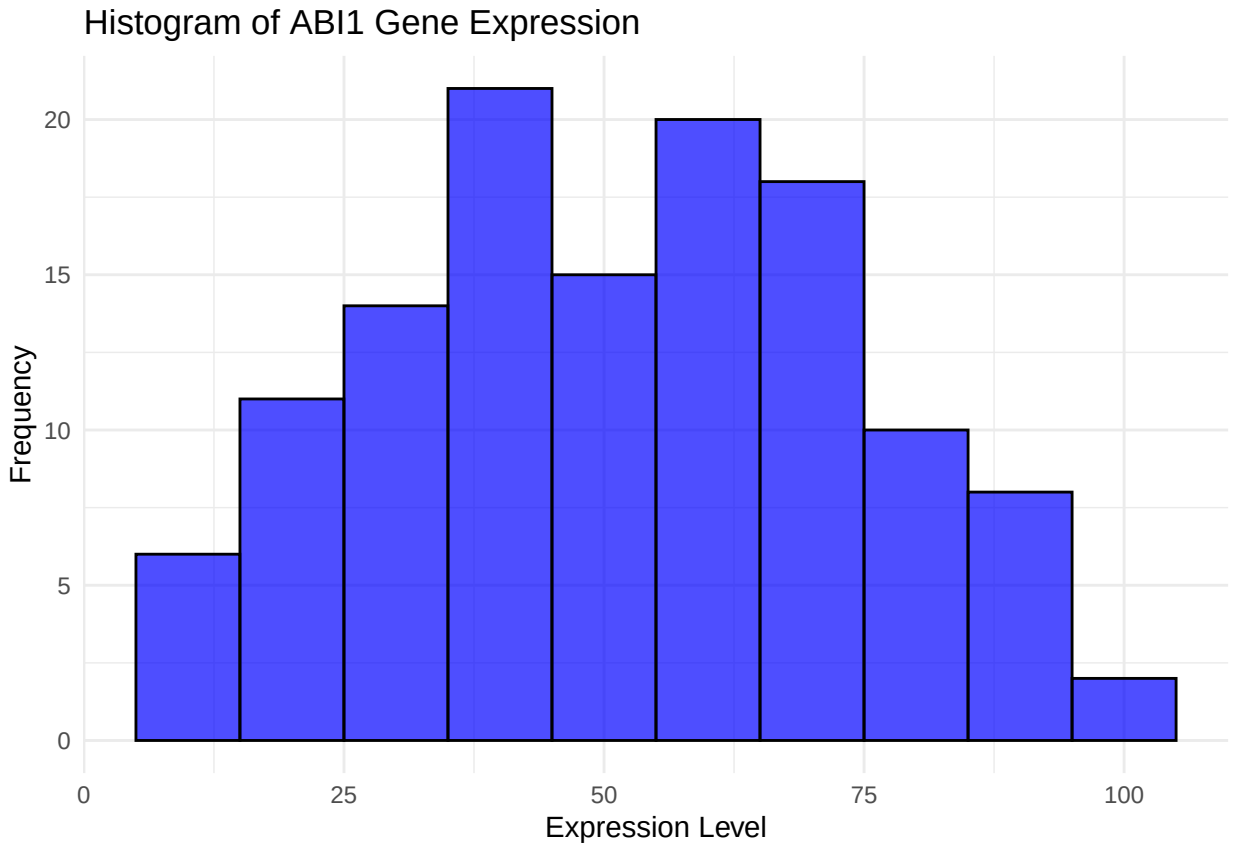
```
      ) %>%
      select(expression, !!sym(continuous_covariate), !!sym(categorical_covariate1), !!sym(categorical_
      distinct()  # Ensure no duplicate rows

    # Check if the gene_data has the correct number of data points
    print(paste("Number of data points for gene", gene, ":", nrow(gene_data)))

     # Histogram: distribution of expression values
    histogram_plot <- ggplot(gene_data, aes(x = expression)) +
      geom_histogram(binwidth = 10, fill = "blue", color = "black", alpha = 0.7) +
      labs(title = paste("Histogram of", gene, "Gene Expression"),
           x = "Expression Level",
           y = "Frequency") +
      theme_minimal()

    print(histogram_plot)

    # Scatterplot: expression vs continuous covariate, colored by categorical covariate1
    scatter_plot <- ggplot(gene_data, aes_string(x = continuous_covariate, y = "expression", color = ca
      geom_point(size = 3, alpha = 0.7) +
      scale_x_continuous(breaks = seq(0, 100, by = 10)) +  # Customize x-axis breaks if needed
      labs(title = paste("Scatterplot of", gene, "Expression vs", continuous_covariate),
           x = continuous_covariate,
           y = "Expression Level",
           color = categorical_covariate1) +
      theme_minimal()

    print(scatter_plot)

    # Boxplot: expression by categorical_covariate1, separated by categorical_covariate2
    box_plot <- ggplot(gene_data, aes_string(x = categorical_covariate1, y = "expression", fill = categ
      geom_boxplot() +
      labs(title = paste("Boxplot of", gene, "Expression by", categorical_covariate1, "and", categorical
           x = categorical_covariate1,
           y = "Expression Level",
           fill = categorical_covariate2) +
      theme_minimal() +
      facet_wrap(as.formula(paste("~", categorical_covariate2))) +
      theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))

    print(box_plot)


  }
}
```

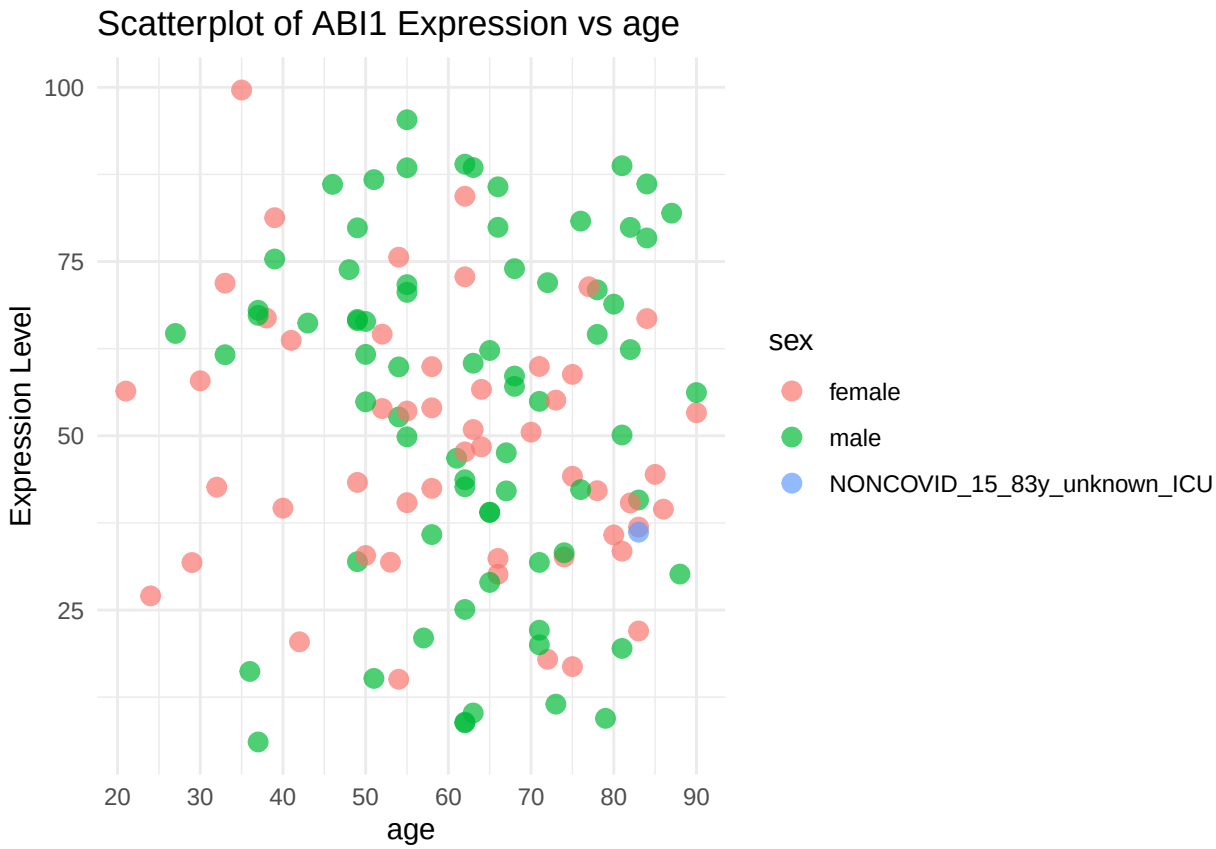
```
# how to use function with three genes
generate_plots(
  data = combined_data,
  genes = c("ABI1", "ABHD17A", "ABAT"),
  continuous_covariate = "age",
  categorical_covariate1 = "sex",
  categorical_covariate2 = "mechanical_ventilation"
```

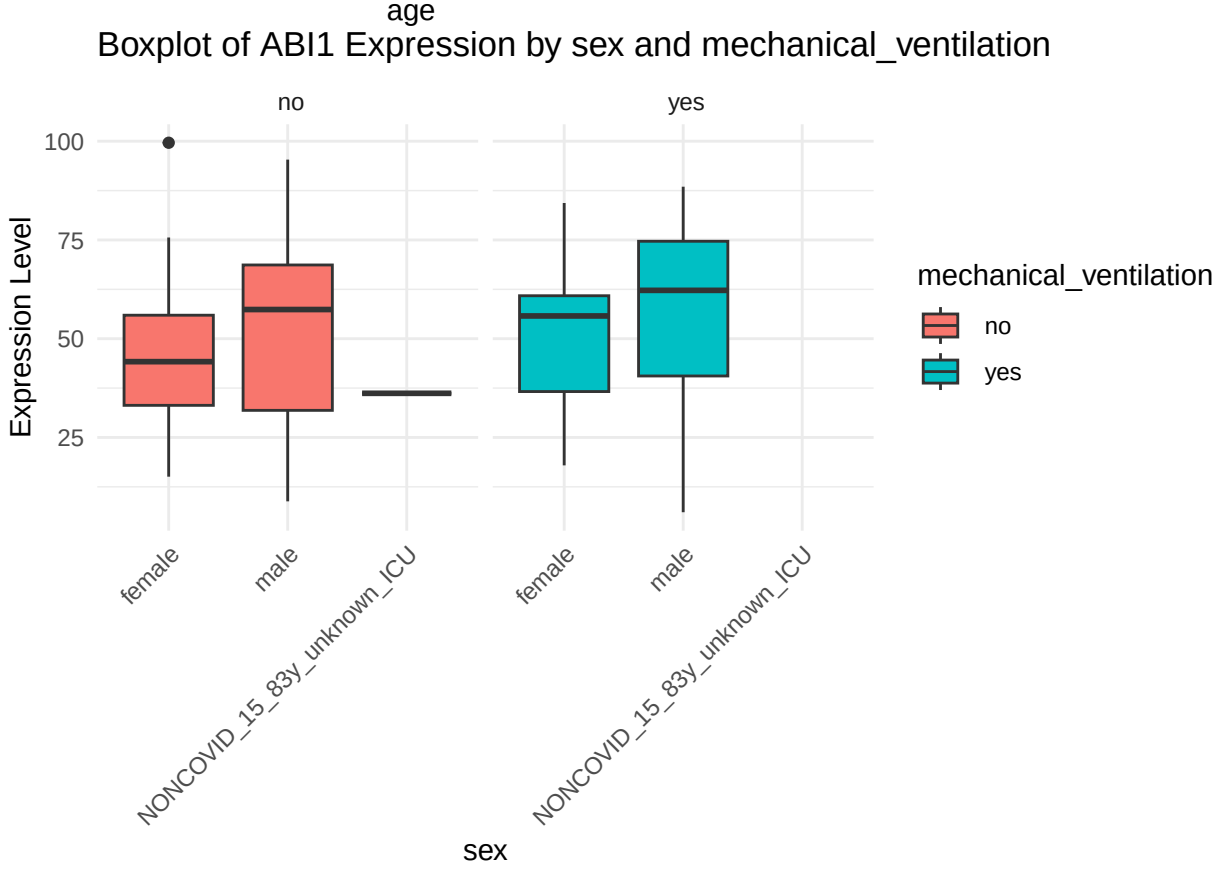
```
)
```

```
## [1] "Number of data points for gene ABI1 : 125"
```

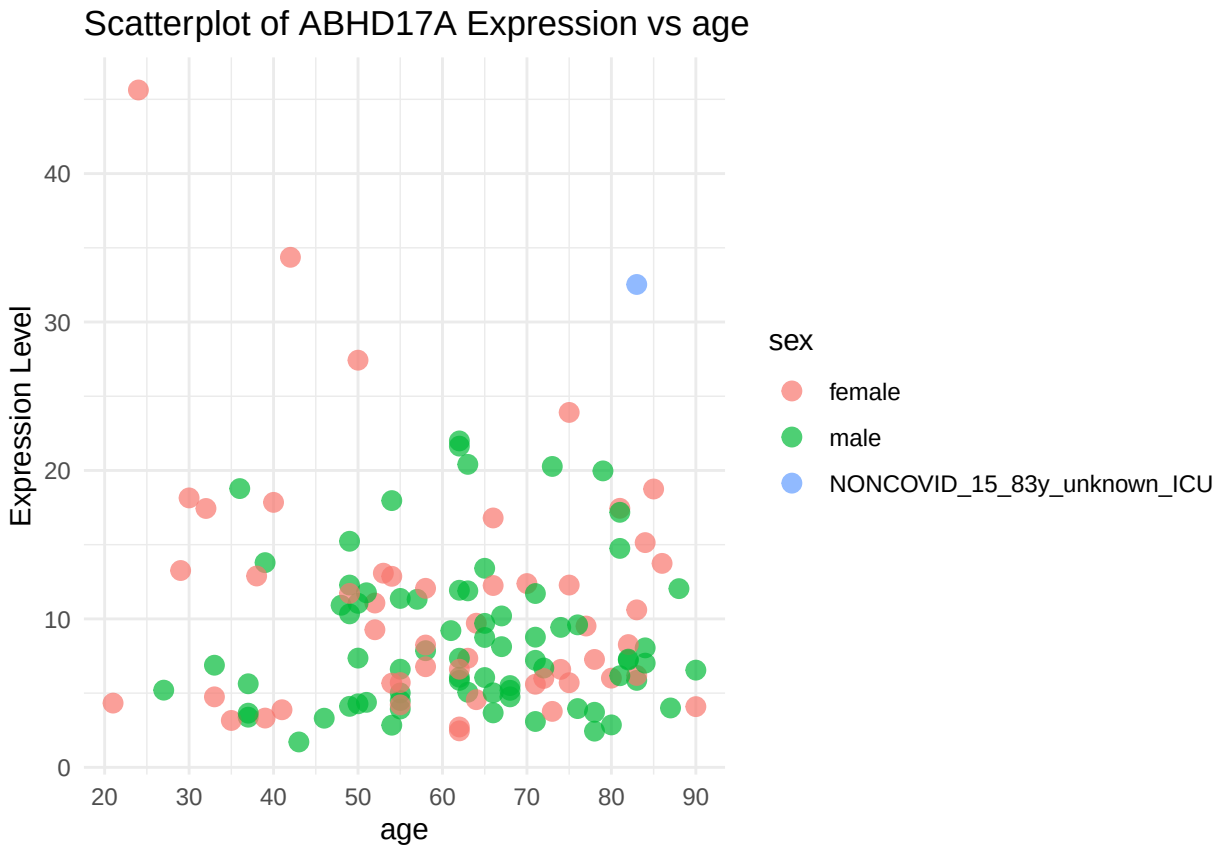
```
## Warning: ‘aes_string()‘ was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with ‘aes()‘.
## i See also ‘vignette("ggplot2-in-packages")‘ for more information.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

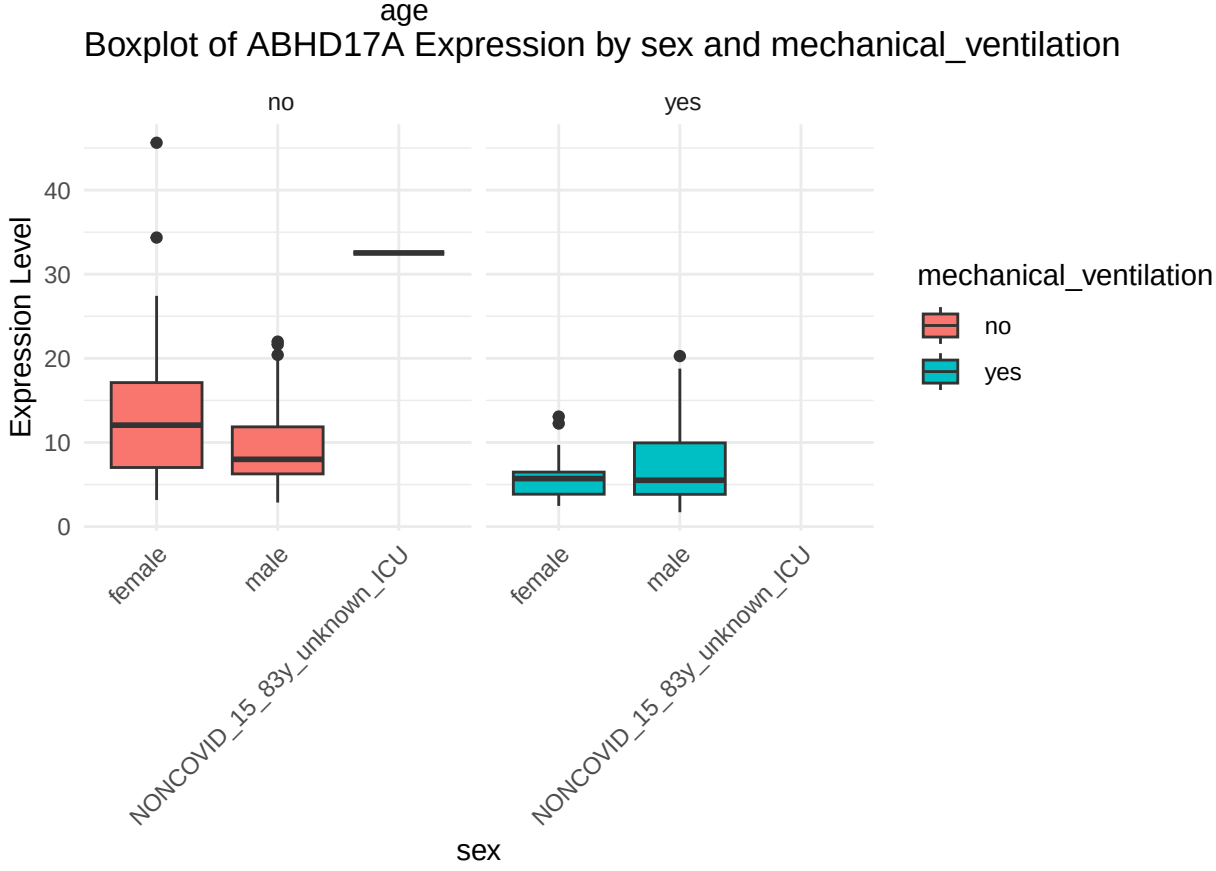



Scatterplot of ABI1 Expression vs age
Expression Level
100
75
50
25
20
30
40
50
60
70
80
90
age
sex
female
male
NONCOVID_15_83y_unknown_ICU

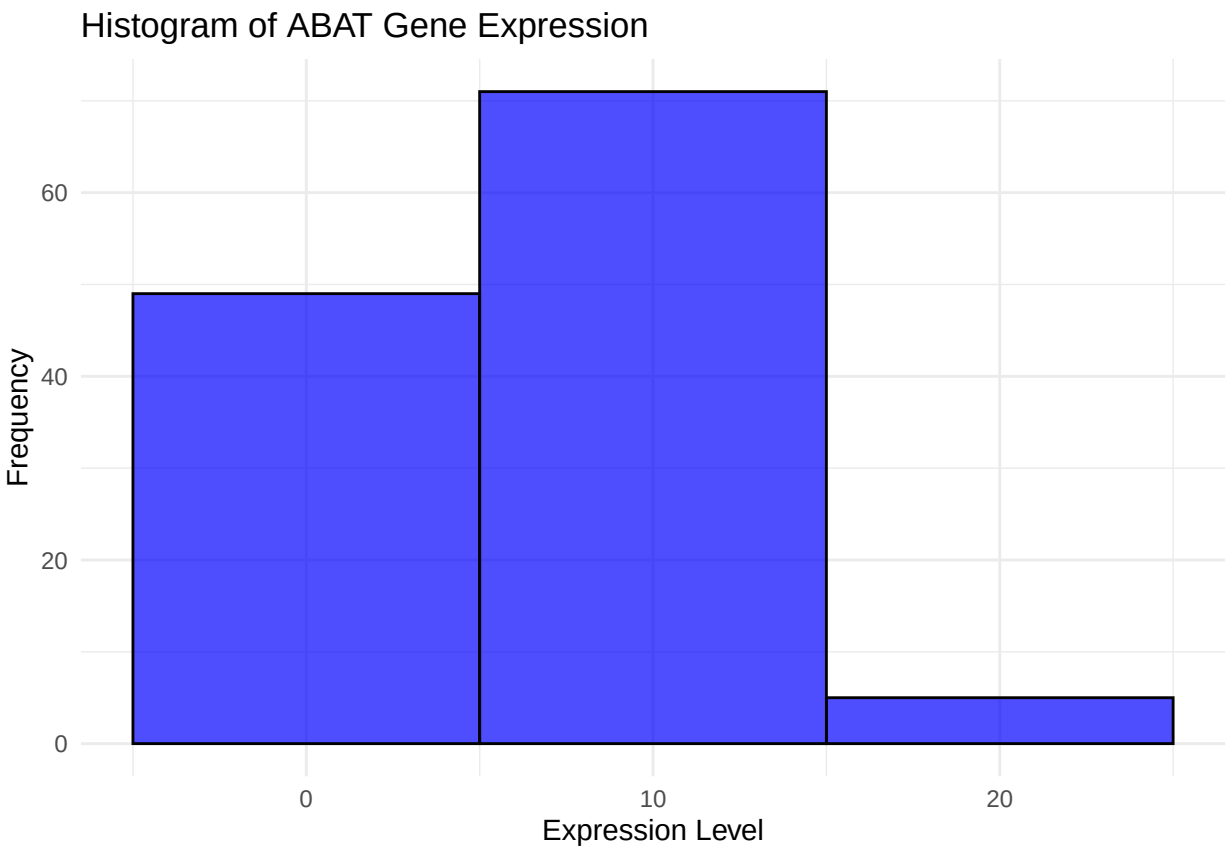

Boxplot of ABI1 Expression by sex and mechanical_ventilation
no
yes
Expression Level
100
75
50
25
female
male
NONCOVID_15_83y_unknown_ICU
female
male
NONCOVID_15_83y_unknown_ICU
sex
mechanical_ventilation
no
yes

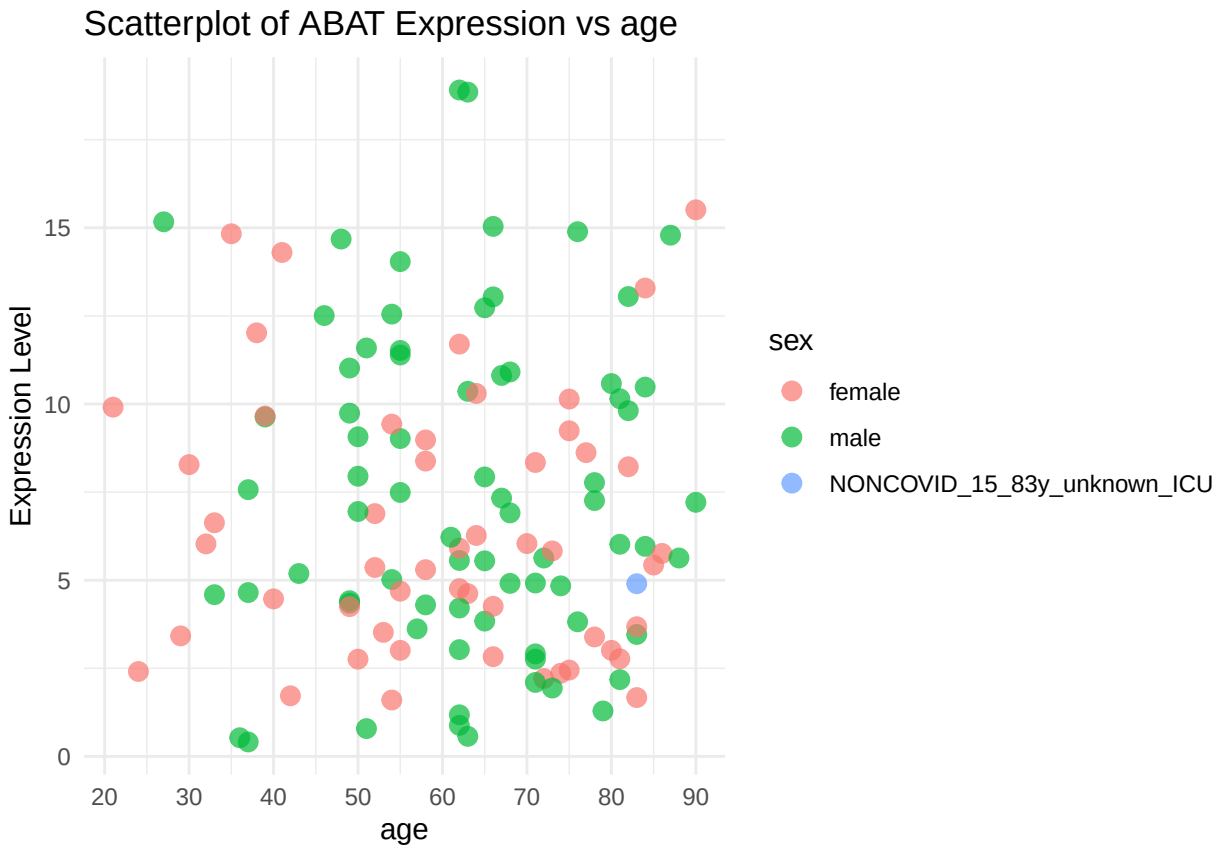
```
## [1] "Number of data points for gene ABHD17A : 125"
```

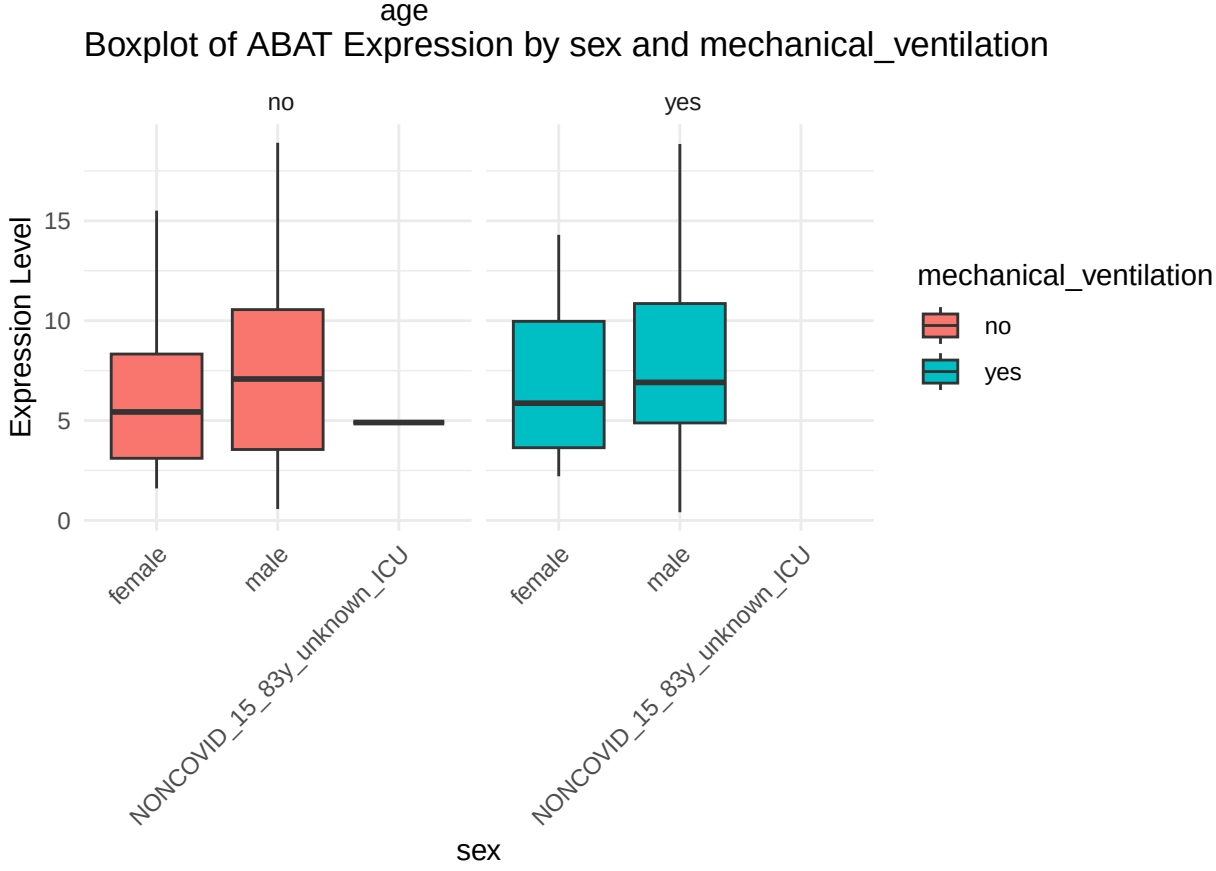

Scatterplot of ABHD17A Expression vs age
Expression Level
40
30
20
10
0
20
30
40
50
60
70
80
90
age
sex
female
male
NONCOVID_15_83y_unknown_ICU


Boxplot of ABHD17A Expression by sex and mechanical_ventilation
no
yes
Expression Level
40
30
20
10
0
female
male
NONCOVID_15_83y_unknown_ICU
female
male
NONCOVID_15_83y_unknown_ICU
sex
mechanical_ventilation
no
yes

```
## [1] "Number of data points for gene ABAT : 125"
```




Scatterplot of ABAT Expression vs age
Expression Level
15
10
5
0
20
30
40
50
60
70
80
90
age
sex
female
male
NONCOVID_15_83y_unknown_ICU


Boxplot of ABAT Expression by sex and mechanical_ventilation
no
yes
Expression Level
15
10
5
0
female
male
NONCOVID_15_83y_unknown_ICU
female
male
NONCOVID_15_83y_unknown_ICU
sex
mechanical_ventilation
no
yes