Final Project Checkpoint 2

John DeForest

July 29, 2024

```r
#load df:
g1 = read.csv('C:\\Users\\John DeForest\\Desktop\\qbs103 R DS\\QBS103_GSE157103_genes.csv')
sm1 = read.csv('C:\\Users\\John DeForest\\Desktop\\qbs103 R DS\\QBS103_GSE157103_series_matrix-1.csv')
#gene, metadata
#check df headers:
#print(head(g1))
#assign first column name as "Gene"
colnames(g1)[1] = 'Gene'
#print(head(g1))

#print(head(sm1))
```

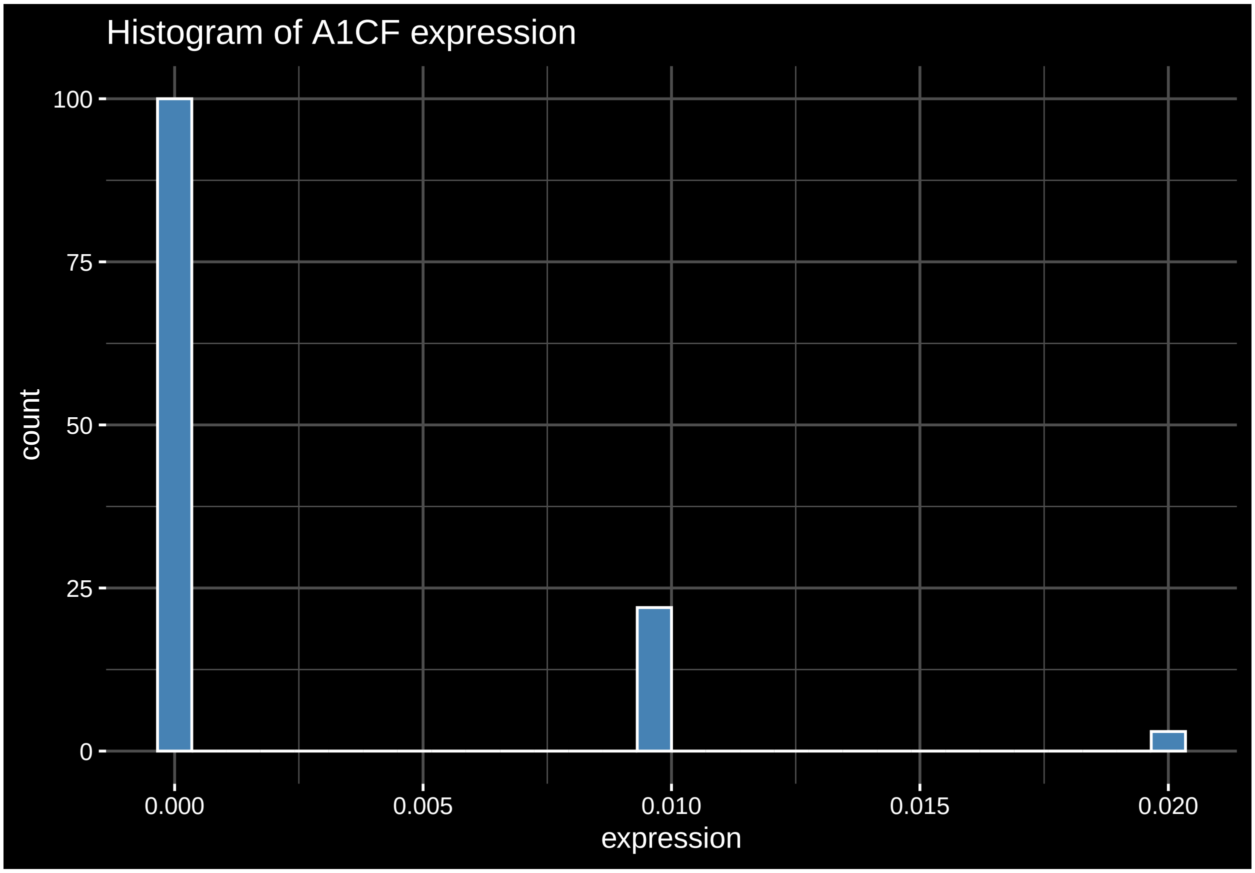
```r
#new function to do full plotting
geneda = function(dfs,genenames,contcovar,catcovar1,catcovar2){
    G1 = dfs[[1]]
    SM1 = dfs[[2]]
    for (g in genenames) {
        #extract gene data from that row
        grow = G1[G1$Gene == g,-1] #select rows where Gene is g
        #print(grow)
        #transpose
        growtp = data.frame(participant_id = colnames(G1)[-1],
                            expression = as.numeric(unlist(grow)))
        #convert to data frame with pid and expression(of gene g)
        #print(head(growtp))
        #add metadata by merge w sm1
        mg = merge(growtp, SM1, by = 'participant_id') #merge by participant_id
        #print(head(mg))

        #fix contcovar if not truly continuous (like age w '>89' as a val lol)
        if (contcovar %in% colnames(mg)) {
  mg[[contcovar]] <- as.character(mg[[contcovar]])

  # Replace ">89" with "90", or any similar fixes
  mg[[contcovar]][mg[[contcovar]] == ">89"] <- "90"

  # Coerce to numeric
  mg[[contcovar]] <- as.numeric(mg[[contcovar]])
}
        #make sure categorical covariates are factors
        mg[[catcovar2]] <- as.factor(mg[[catcovar2]])
        mg[[catcovar1]] <- as.factor(mg[[catcovar1]])
    #1 histogram of gene expression
    library(ggplot2)
    #summary(mg$expression)
    #hist(mg$expression) #duh
    histoplot <- ggplot(mg, aes(x = expression)) +
    geom_histogram(color = "white", fill = "steelblue", bins = 30) +
    ggtitle(paste("Histogram of", g, "expression")) +
    theme_minimal()+
theme(
    plot.title = element_text(color = "white"),
    axis.title = element_text(color = "white"),
    axis.text = element_text(color = "white"),
    axis.ticks = element_line(color = "white"),
    panel.grid = element_line(color = "gray30"),
    panel.background = element_rect(fill = "black"),
    plot.background = element_rect(fill = "black"),
    panel.border = element_blank()
)
    # Save the plot
    hist_file <- paste0(g, '_exp_hist.png',sep='')
    ggsave(hist_file, plot = histoplot, width = 8, height = 6)

    #2 Scatterplot of Gene Expression and Age (contcovar)
    scatplot = ggplot(mg, aes_string(x = contcovar, y = "expression")) +
        geom_point(color = "white") +
        labs(title = paste("Scatterplot of", g, "Expression vs", contcovar),
             x = paste(contcovar, "(years)"),
             y = paste(g, "Expression ")) +
        theme_minimal()+
theme(
    plot.title = element_text(color = "white"),
    axis.title = element_text(color = "white"),
    axis.text = element_text(color = "white"),
    axis.ticks = element_line(color = "white"),
    panel.grid = element_line(color = "gray30"),
    panel.background = element_rect(fill = "black"),
    plot.background = element_rect(fill = "black"),
    panel.border = element_blank()
)
    scat_file <- paste0(g, '_exp_vs_',contcovar,'_scat.png',sep='')
    ggsave(scat_file, plot = scatplot, width = 8, height = 6)

    #3 Boxplot of Gene Expression by Categorical Covariates (ie Sex and mechanical_ventilation)
    title1 = paste("Boxplot of", g, "Expression by", catcovar1,'Colored by',catcovar2)
    boxo1 = ggplot(mg, aes_string(x = catcovar1, y = "expression", fill = catcovar2)) +
        geom_boxplot(color = "white") +
        labs(title = title1,
             x = catcovar1,
             y = paste(g, "Expression "),
             fill = paste(catcovar2)
         ) +
        theme_minimal()+
theme(
        plot.title = element_text(color = "white"),
        axis.title = element_text(color = "white"),
        axis.text = element_text(color = "white"),
        axis.ticks = element_line(color = "white"),
        panel.grid = element_line(color = "gray30"),
        panel.background = element_rect(fill = "black"),
        plot.background = element_rect(fill = "black"),
        panel.border = element_blank(),
        legend.title = element_text(color = "white"),
legend.text = element_text(color = "white")

    )
        box1_file <- paste0(g, '_box_exp_by_',catcovar1,'_col_by_',catcovar2,'.png',sep='')
        ggsave(box1_file, plot = boxo1, width = 8, height = 6)

        title2 = paste("Boxplot of", g, "Expression by", catcovar2,'Colored by',catcovar1)
        boxo2 = ggplot(mg, aes_string(x = catcovar2, y = "expression", fill = catcovar1)) +
            geom_boxplot(color = "white") +
            labs(title = title2,
                 x = catcovar2,
                 y = paste(g, "Expression "),
                 fill = paste(catcovar1)
             ) +
            theme_minimal()+
    theme(
        plot.title = element_text(color = "white"),
        axis.title = element_text(color = "white"),
        axis.text = element_text(color = "white"),
        axis.ticks = element_line(color = "white"),
        panel.grid = element_line(color = "gray30"),
        panel.background = element_rect(fill = "black"),
        plot.background = element_rect(fill = "black"),
        panel.border = element_blank(),
        legend.title = element_text(color = "white"),
legend.text = element_text(color = "white")

    )
        box2_file <- paste0(g, '_box_exp_by_',catcovar2,'_col_by_',catcovar1,'.png',sep='')
        ggsave(box2_file, plot = boxo2, width = 8, height = 6)

        #make sure plots plot in knitting:
        print(histoplot)
        print(scatplot)
        print(boxo1)
        print(boxo2)
    }

}
```

```
#driver code for function call
gene_sel = 'A1CF' #same as FP1 for verification
geneda(dfs = list(g1, sm1),
        genenames = c(gene_sel),
        contcovar = 'age',
        catcovar1 = 'mechanical_ventilation',
        catcovar2 = 'sex')
```

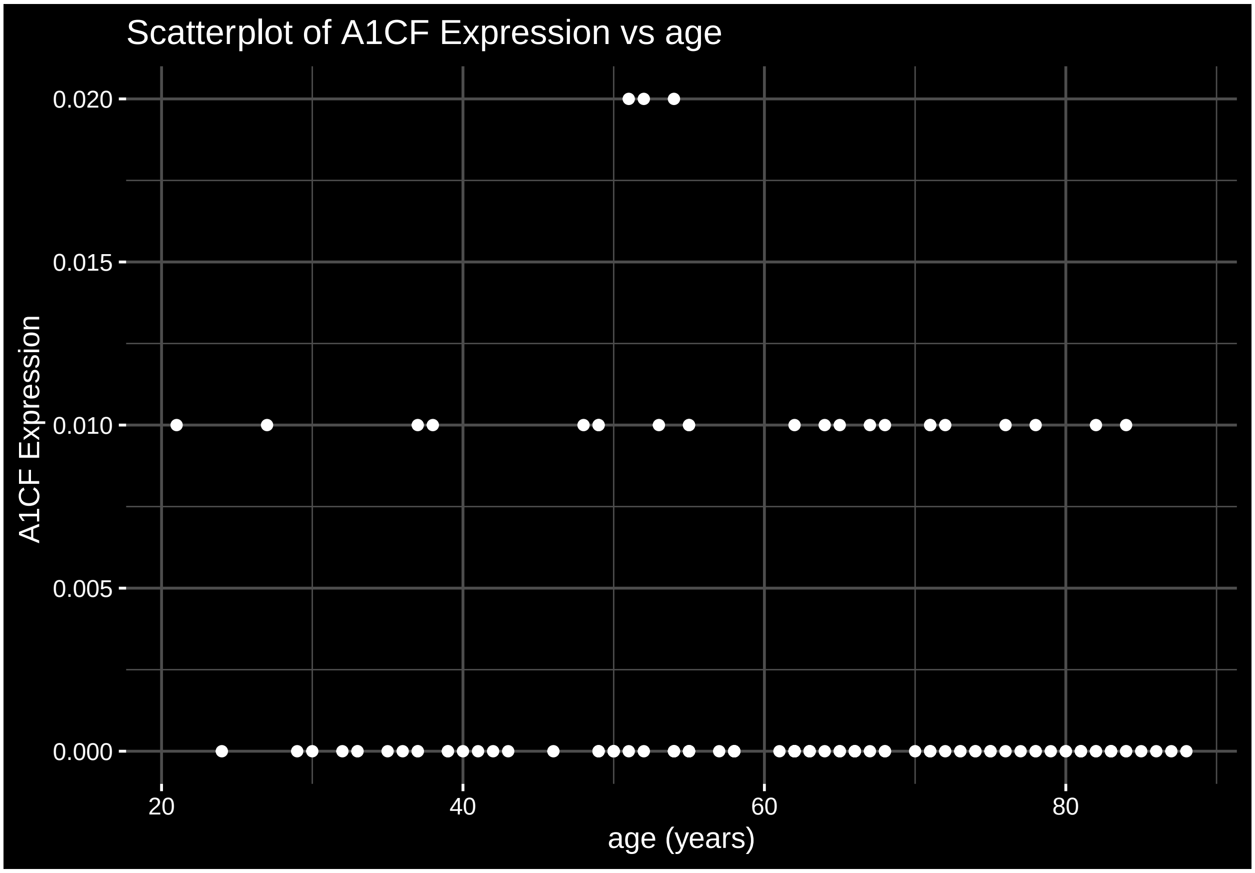
```
## Warning in geneda(dfs = list(g1, sm1), genenames = c(gene_sel), contcovar =
## "age", : NAs introduced by coercion
```

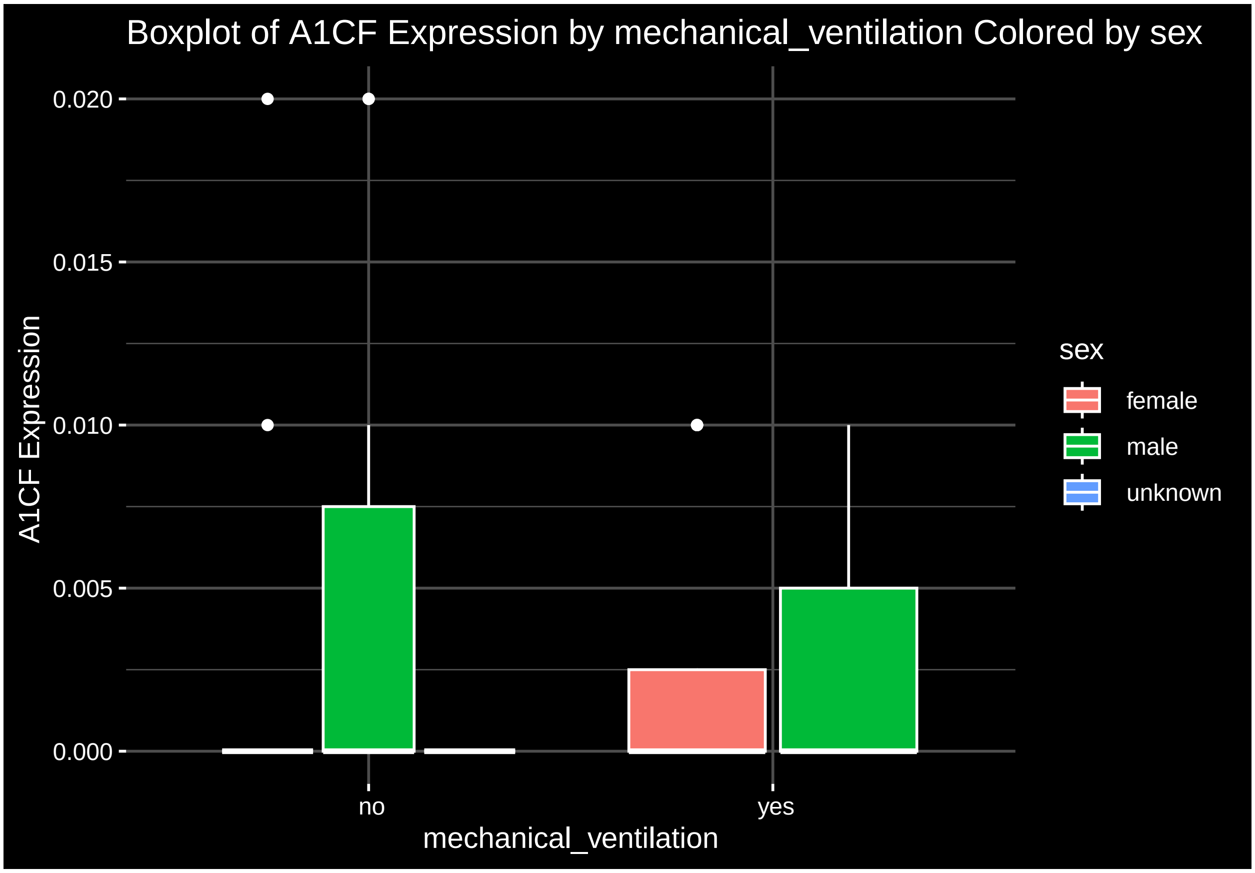
```
## Warning: ‘aes_string()‘ was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with ‘aes()‘.
## i See also ‘vignette("ggplot2-in-packages")‘ for more information.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this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

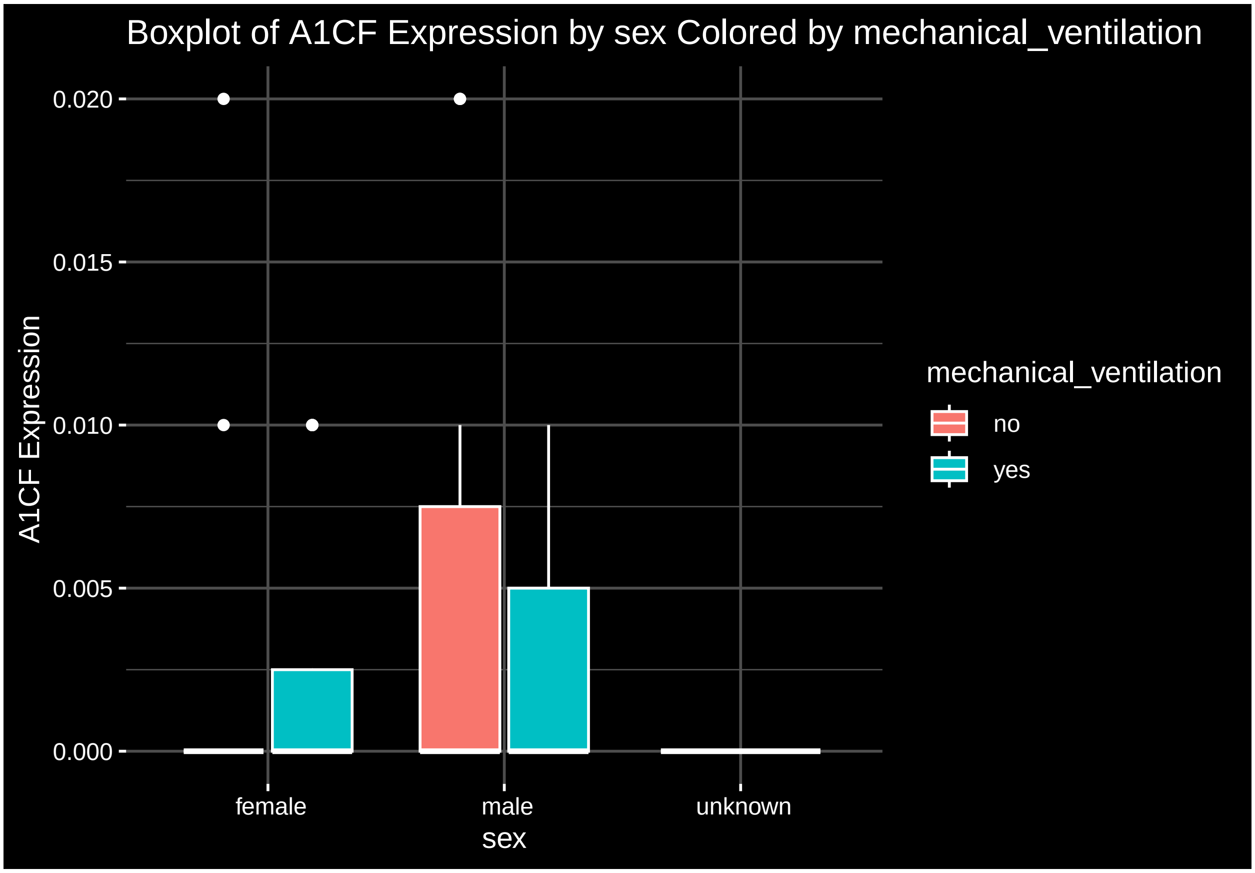
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```


Scatterplot of A1CF Expression vs age
0.020
0.015
0.010
0.005
0.000
A1CF Expression
20
40
60
80
age (years)

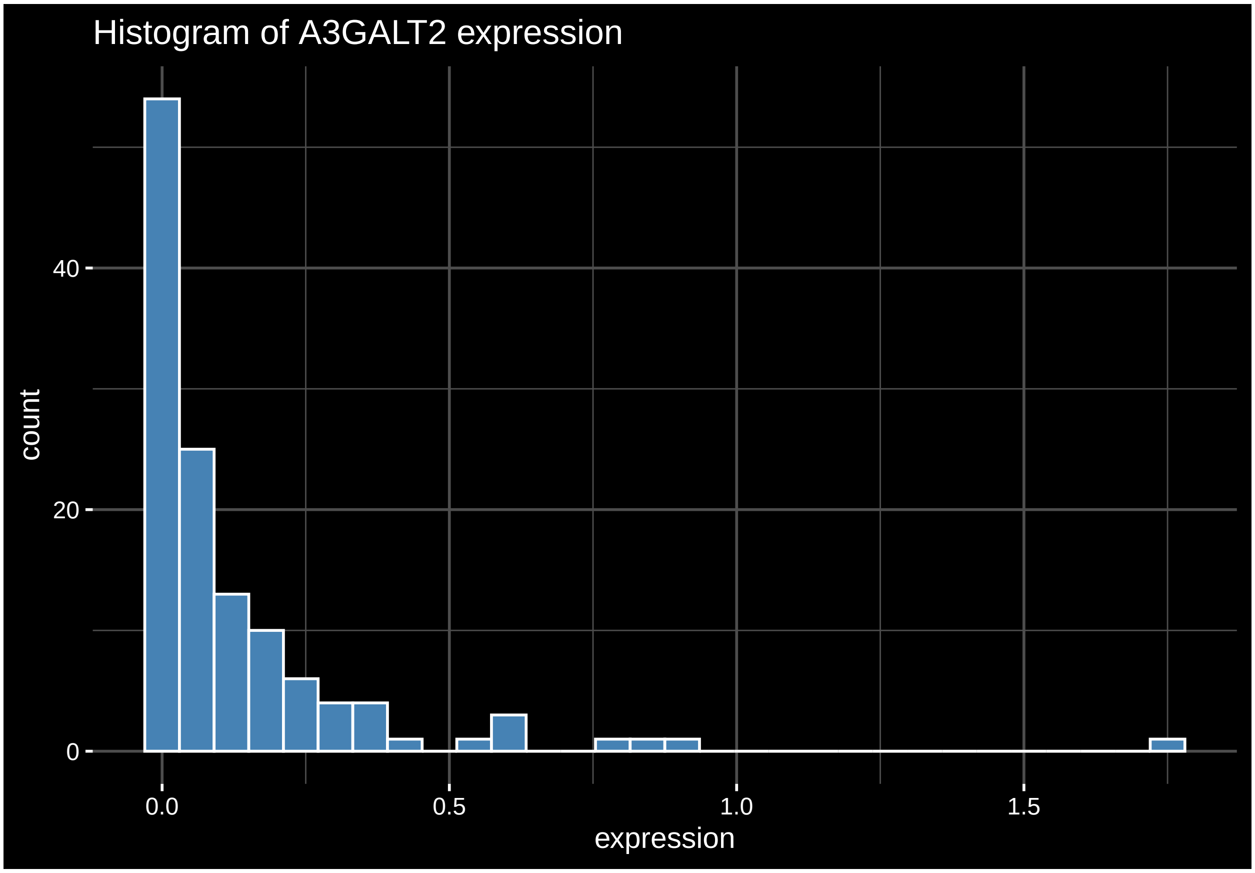

Boxplot of A1CF Expression by mechanical_ventilation Colored by sex
A1CF Expression
0.020
0.015
0.010
0.005
0.000
no
yes
mechanical_ventilation
sex
female
male
unknown



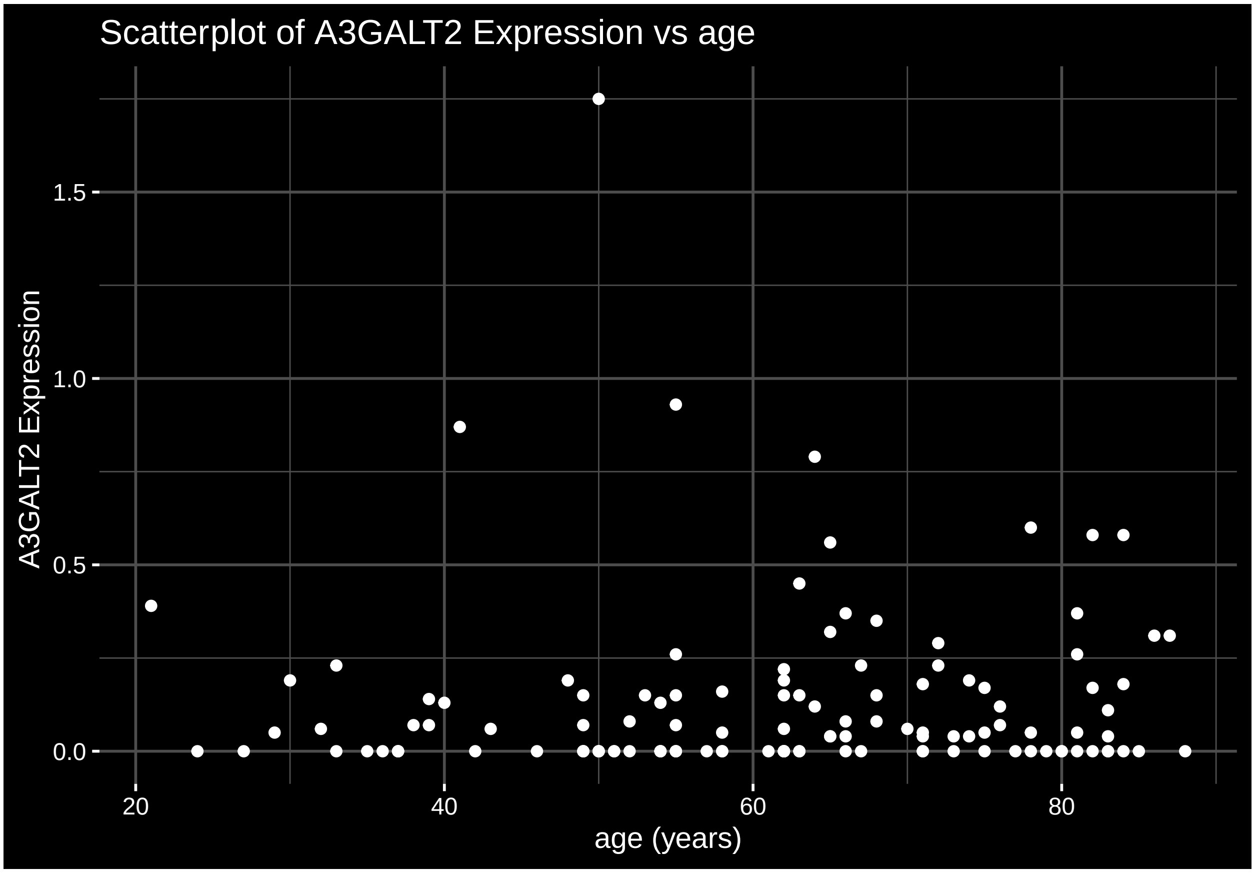
```r
#Now 2 Additional Genes:
moregenes = c('A3GALT2','A4GALT')
#print(head(g1))

for (g in moregenes) {
    geneda(dfs = list(g1, sm1),
        genenames = c(g),
        contcovar = 'age',
        catcovar1 = 'mechanical_ventilation',
        catcovar2 = 'sex')
    cat("###", g, "\n\n")
    knitr::include_graphics(paste0(g, "_exp_hist.png",sep=''))
    knitr::include_graphics(paste0(g, "_exp_vs_age_scat.png",sep=''))
    knitr::include_graphics(paste0(g, "_box_exp_by_mechanical_ventilation_col_by_sex.png",sep=''))
    knitr::include_graphics(paste0(g, "_box_exp_by_sex_col_by_mechanical_ventilation.png",sep=''))

}
```

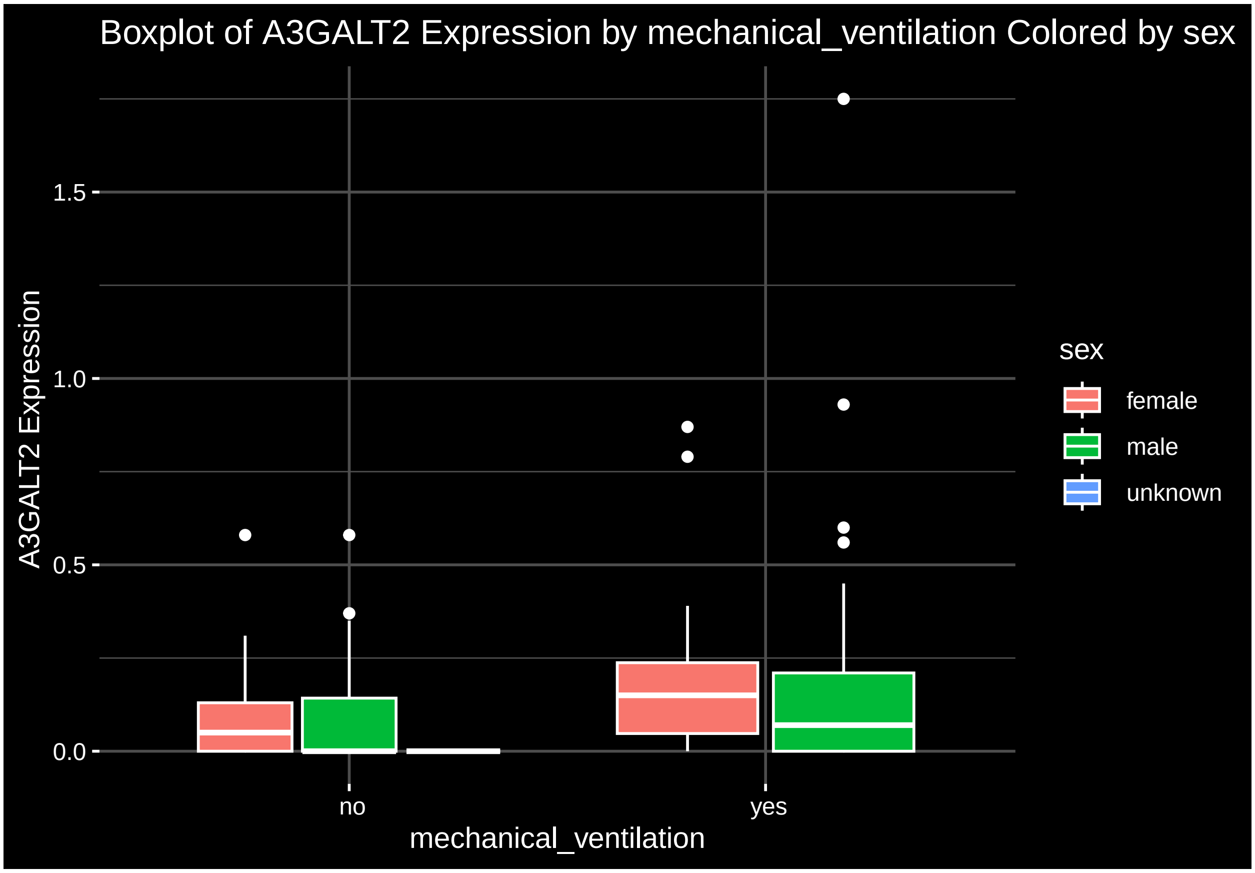
```
## Warning in geneda(dfs = list(g1, sm1), genenames = c(g), contcovar = "age", :
## NAs introduced by coercion
```

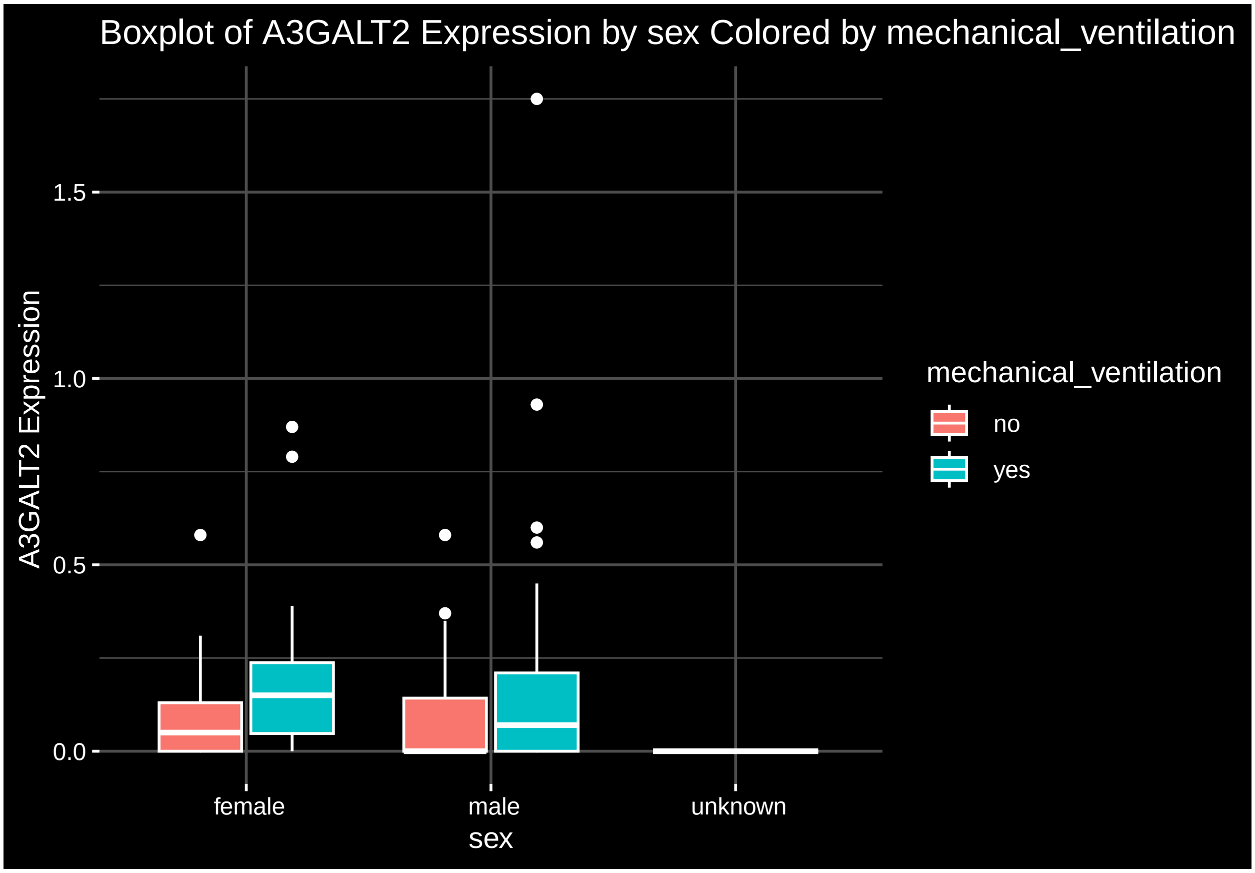
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



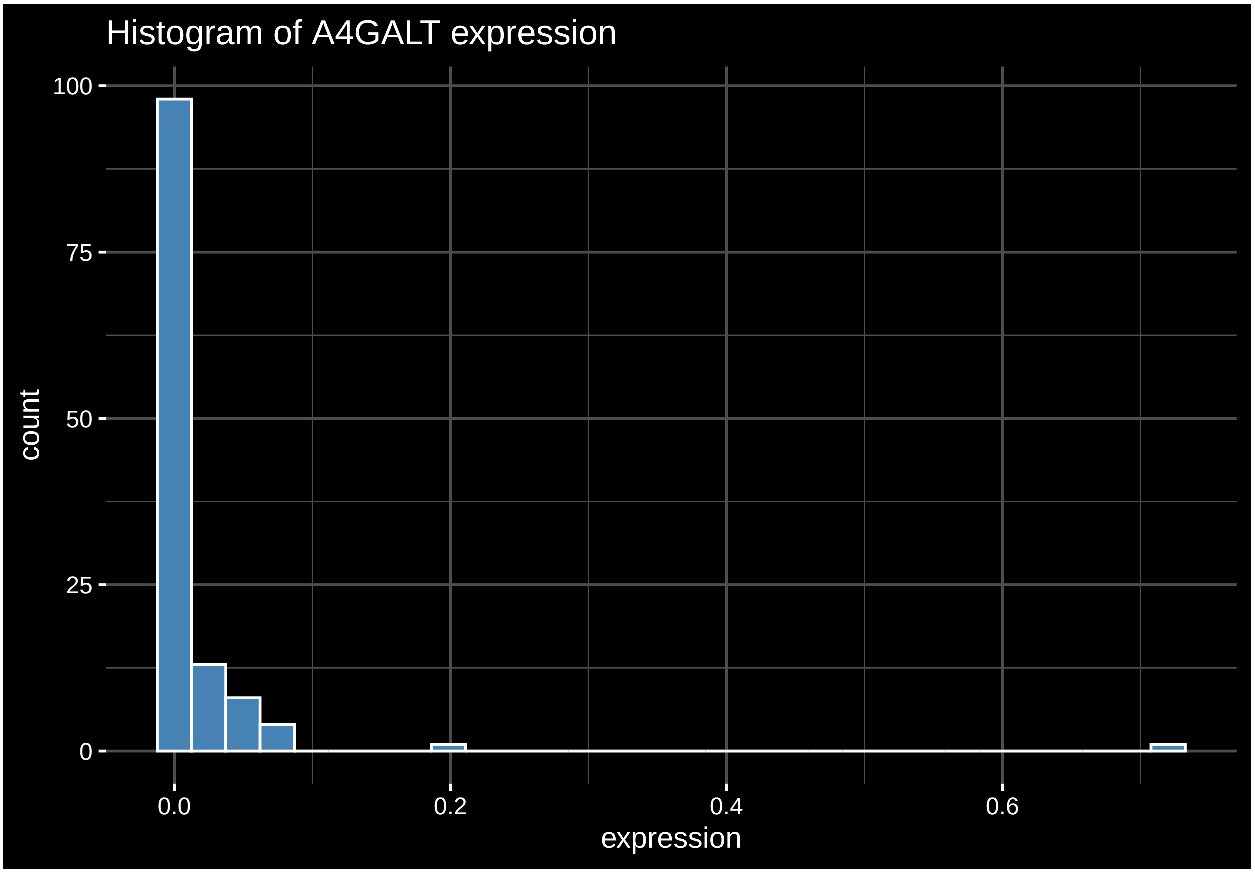
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```


Scatterplot of A3GALT2 Expression vs age
A3GALT2 Expression
0.0
0.5
1.0
1.5
20
40
60
80
age (years)

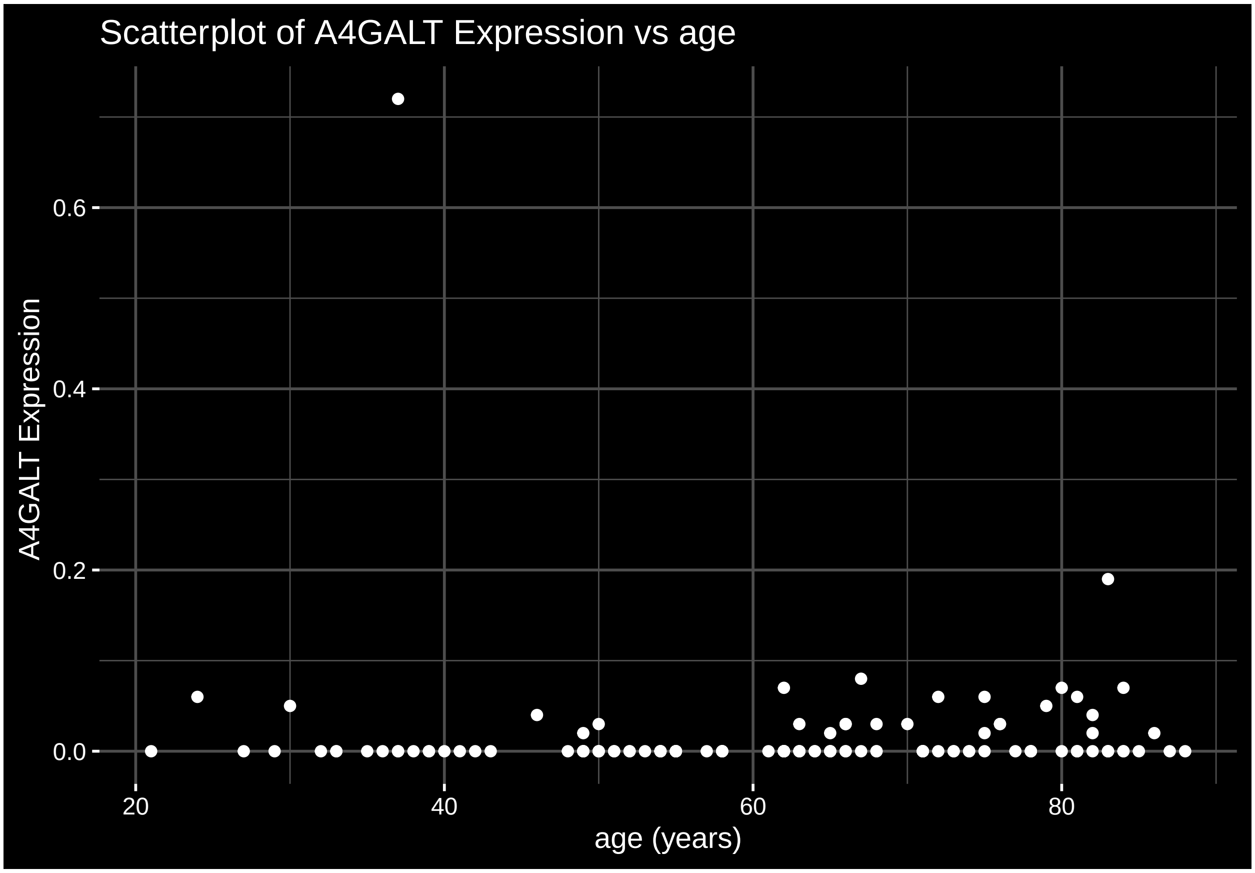

Boxplot of A3GALT2 Expression by mechanical_ventilation Colored by sex
A3GALT2 Expression
1.5
1.0
0.5
0.0
no
yes
mechanical_ventilation
sex
female
male
unknown

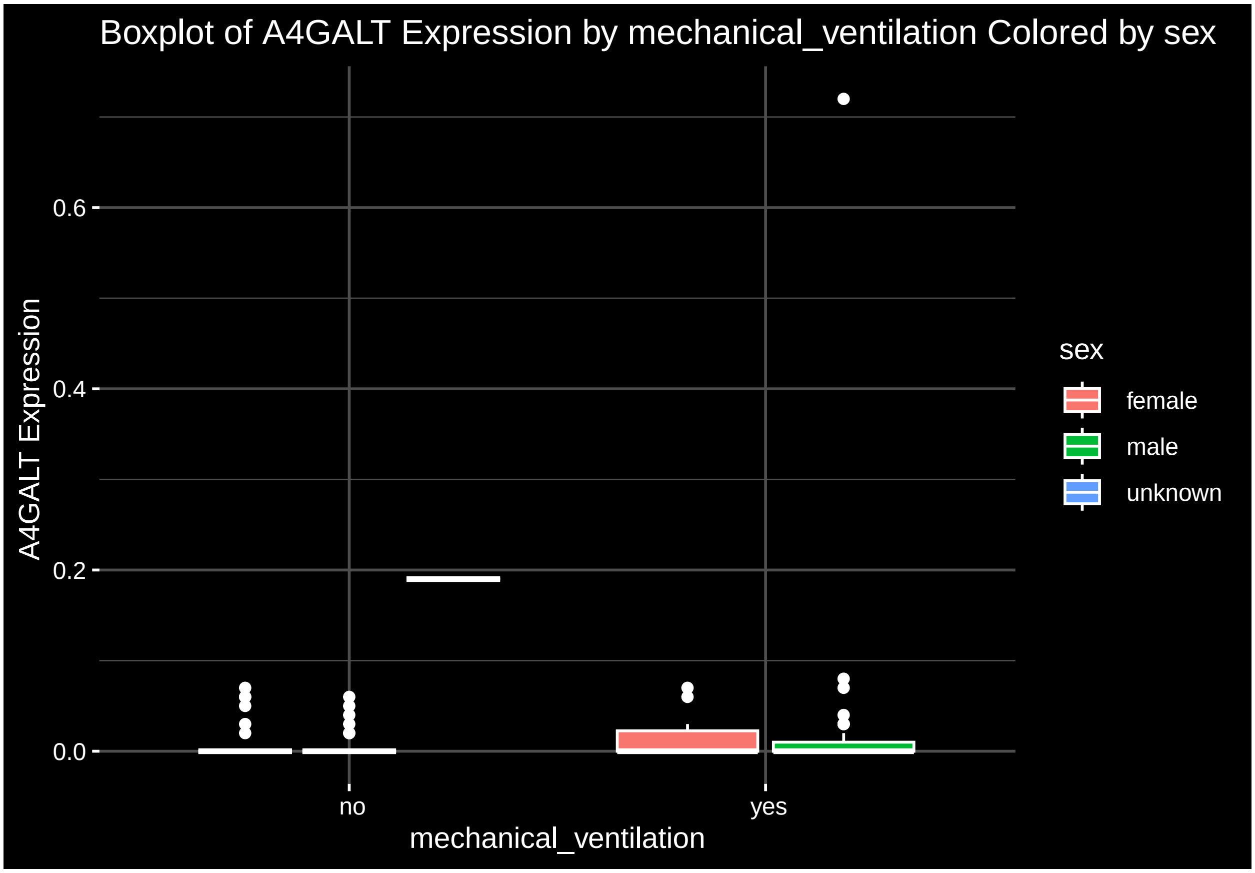


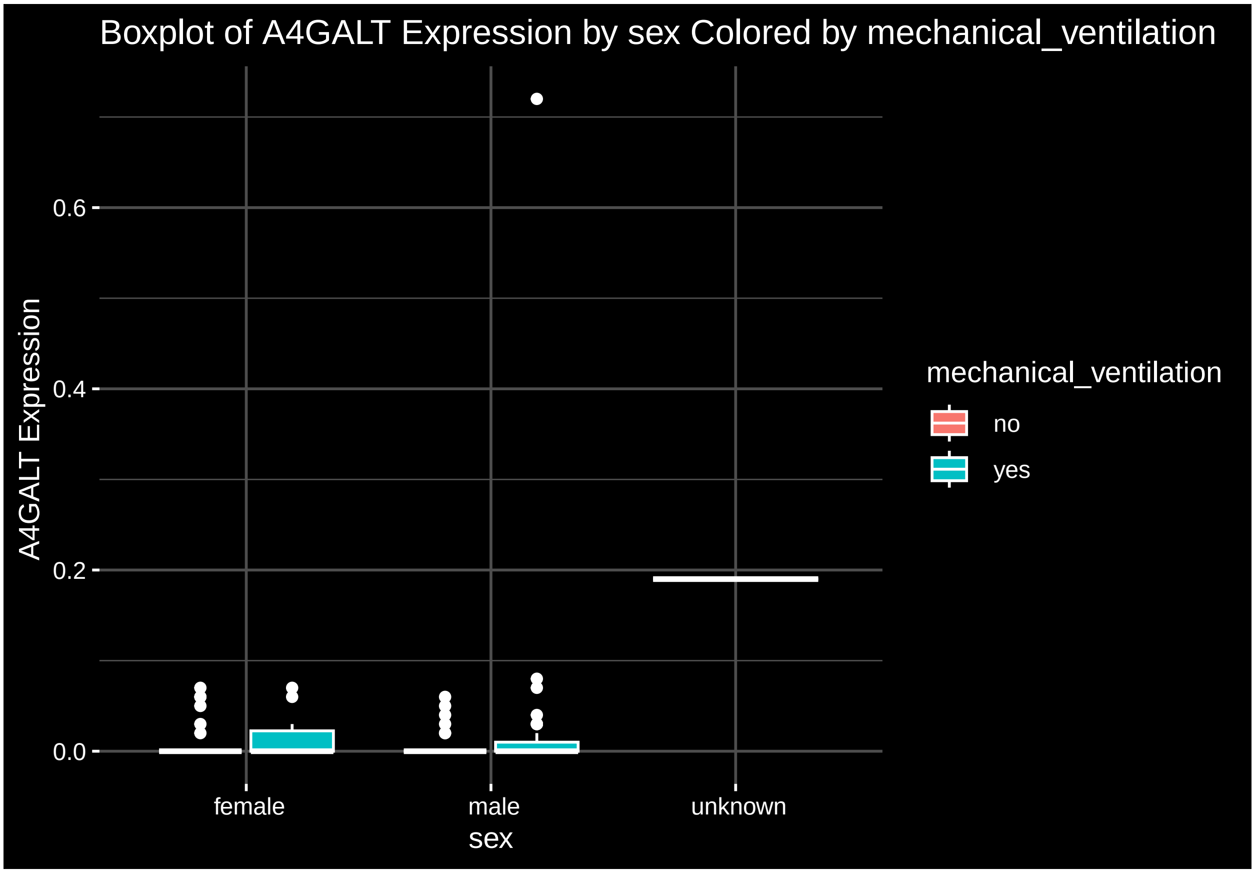
```
## ### A3GALT2

## Warning in geneda(dfs = list(g1, sm1), genenames = c(g), contcovar = "age", : NAs introduced by coerc
## Warning in geneda(dfs = list(g1, sm1), genenames = c(g), contcovar = "age", : Removed 2 rows contain
## (‘geom_point()‘).
```



```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```


Scatterplot of A4GALT Expression vs age
A4GALT Expression
0.0
0.2
0.4
0.6
age (years)
20
40
60
80


Boxplot of A4GALT Expression by mechanical_ventilation Colored by sex
A4GALT Expression
0.0
0.2
0.4
0.6
no
yes
mechanical_ventilation
sex
female
male
unknown



```
## ### A4GALT
```

End of part 2. knit boio