QBS 103 Final Project Part 2

Chichi Illoh

2023-08-08

Final Project Part 2

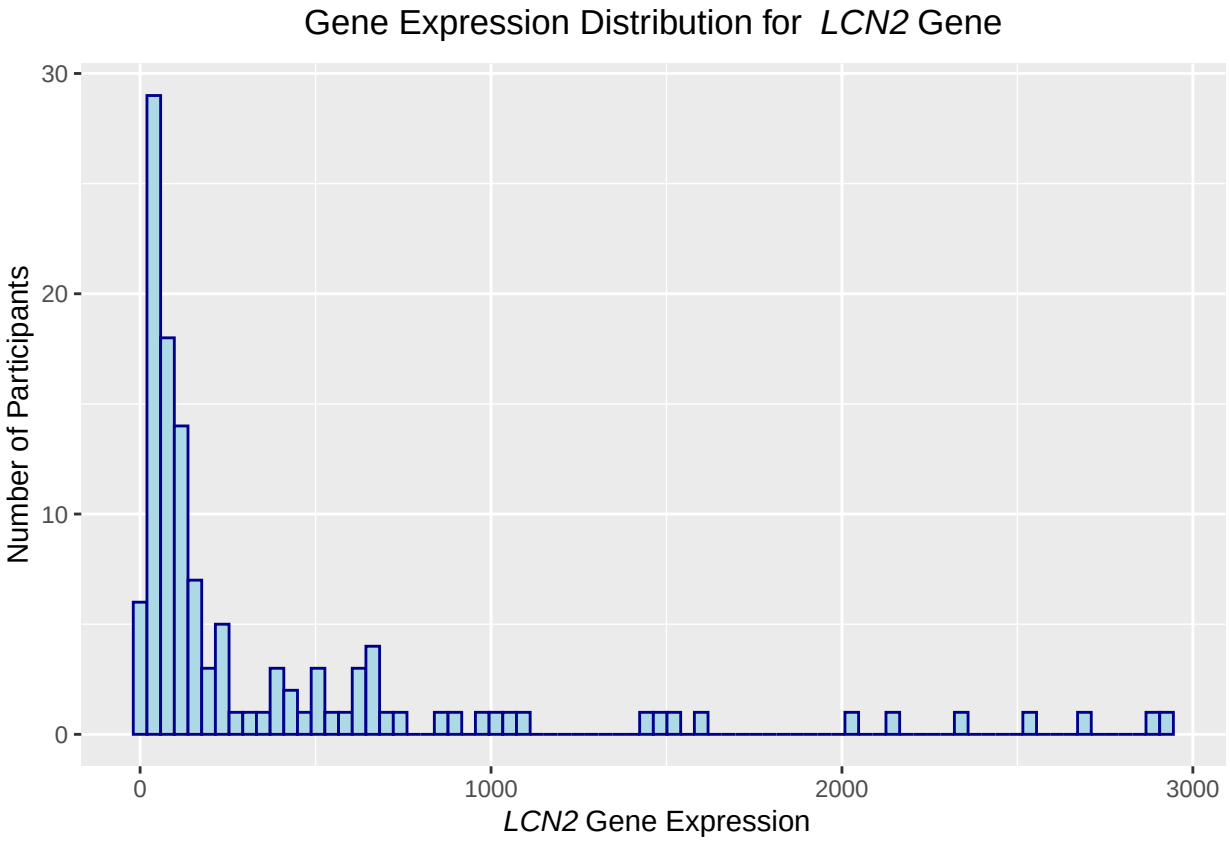
```
meta_data = read.csv("QBS103_finalProject_metadata.csv")
#head(meta_data)
gene_data = read.csv("QBS103_finalProject_geneExpression.csv")
#head(gene_data)
```

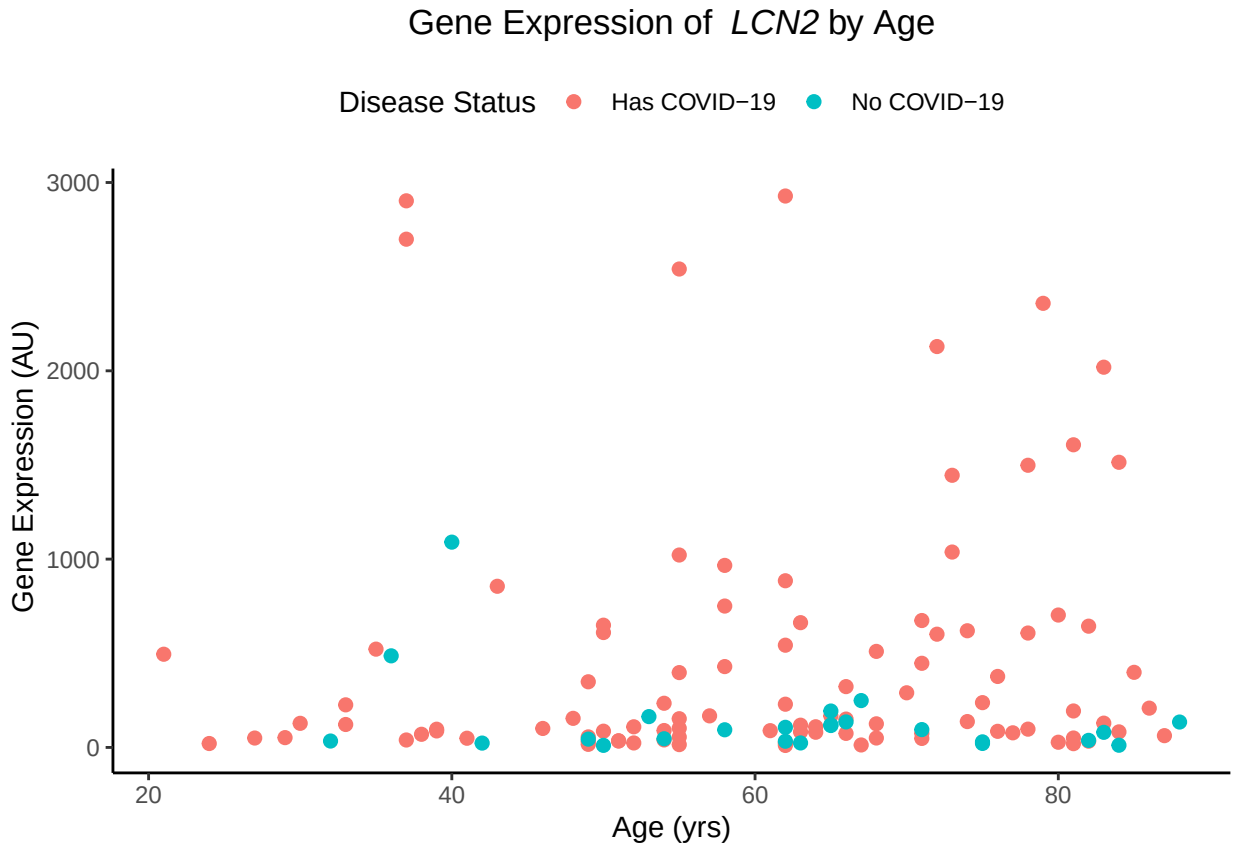
```
stats_plts <- function(data_frame, gene_names,contin_covariate,categorical_var1,categorical_var2)
  {
  # reading the data
  library(tidyverse)
  for (gene in gene_names){
  # convert into vertical data and extract specific gene from gene_names list
    i.gene <- c(gene)
    hist_labels <- c("Gene Expression Distribution for ", "Gene")
    hist_xlabels <- c("Gene Expression")
    hist_title<- eval(bquote(expression(.(hist_labels[1])~italic(.(i.gene[1]))~. (hist_labels[2]))))
    hist_xaxis<-eval(bquote(expression(~italic(.(i.gene[1]))~. (hist_xlabels[1]))))
    scatter_labels <-c("Gene Expression of ", "by Age")
    scatter_title <- eval(bquote(expression(.(scatter_labels[1])~italic(.(i.gene[1]))~. (scatter_labels
    i = which(data_frame$X == gene)
    gene_expression = matrix(data_frame[i,2:127])
    head(gene_expression)

    # new data frame created to link variables from both data sets into one
    newData <- data.frame('Subject ID' = seq(1:126), "Gene.Expression" = gene_expression,'Continuous_Va
    head(newData)

#----------MIGHT HAVE TO REMOVE BELOW-----------------------
    newData <- newData[!(row.names(newData) %in% c("6","86","104","115")),] # filter out data that do
    # Source: https://sparkbyexamples.com/r-programming/drop-dataframe-rows-in-r/?expand_article=1

    newData$Continuous_Var<- as.numeric(newData$Continuous_Var)# Convert age variable to numeric
    newData$Gene.Expression<- as.numeric(newData$Gene.Expression) #convert gene expression variable
```

```r
  # Histogram for gene expression
    histogram<- ggplot(newData, aes(x=Gene.Expression)) +

    geom_histogram(binwidth = 39,color = "darkblue", fill = "lightblue")+

    labs(title = hist_title, x = hist_xaxis , y = "Number of Participants")+

      theme(plot.title = element_text(hjust = 0.5))


    BlankTheme <- theme(# Remove borders and grid lines
        panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        # Define my axis
        axis.line = element_line(colour = "black", linewidth = rel(1)),
        # Set plot background to white
        plot.background = element_rect(fill = "white"),
        panel.background = element_blank(),
        legend.key = element_rect(fill = 'white'),
        # Move legend to the top
        legend.position = 'top')
# Scatterplot for gene expression and continuous covariate
    scatterplot<-ggplot(newData, aes(x = Continuous_Var, y = Gene.Expression, color = Categorical_Var1))

      geom_point(size = 2)+

      labs(title = scatter_title, x = "Age (yrs)", y = " Gene Expression (AU)", color = "Disease Status

      scale_color_manual(values = c("#0066CC", "#CCCCFF")) +

      theme(plot.title = element_text(hjust = 0.5))+

      BlankTheme+

      scale_color_discrete(labels=c('Has COVID-19', 'No COVID-19'))




    BlankTheme <- theme(# Remove borders and grid lines
            panel.border = element_blank(), panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            # Define my axis
            axis.line = element_line(colour = "black", linewidth = rel(1)),
            # Set plot background to white
            plot.background = element_rect(fill = "white"),
            panel.background = element_blank(),
            legend.key = element_rect(fill = 'white'),
            # Move legend to the top
            legend.position = 'top')
    # Box plot of gene expression separated by sex and disease status
```

```
    boxplot<-ggplot(newData, aes(x = Categorical_Var1, y = Gene.Expression, fill = Categorical_Var2))+
      geom_boxplot()+



      labs(title = "Gene Expression of gene grouped by Disease Status \n and categorized by ICU Status"
      BlankTheme+
      theme(plot.title = element_text(hjust = 0.5))+
      scale_x_discrete(labels = c("Has COVID-19", "No COVID-19"))+
      scale_fill_discrete(labels = c("Not in the ICU", "In the ICU"))+
      scale_fill_manual(values = c('#00CC66','#6699CC'))


      plot(histogram)
      plot(scatterplot)
      plot(boxplot)
      print(gene)
}
}


stats_plts(gene_data, gene_names = list("LCN2","CD24","BPI"), meta_data$age, meta_data$disease_status,m
```

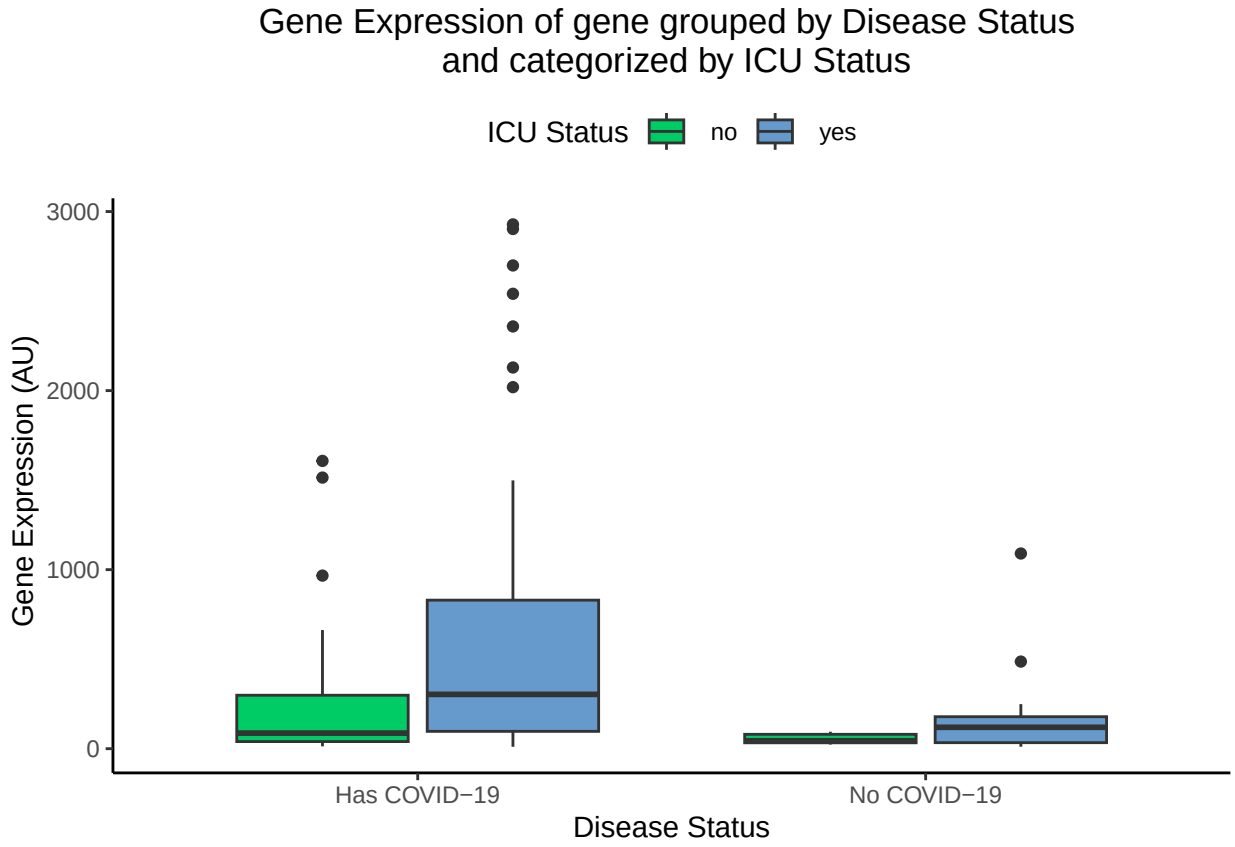
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
## Scale for colour is already present.
## Adding another scale for colour, which will replace the existing scale.
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```

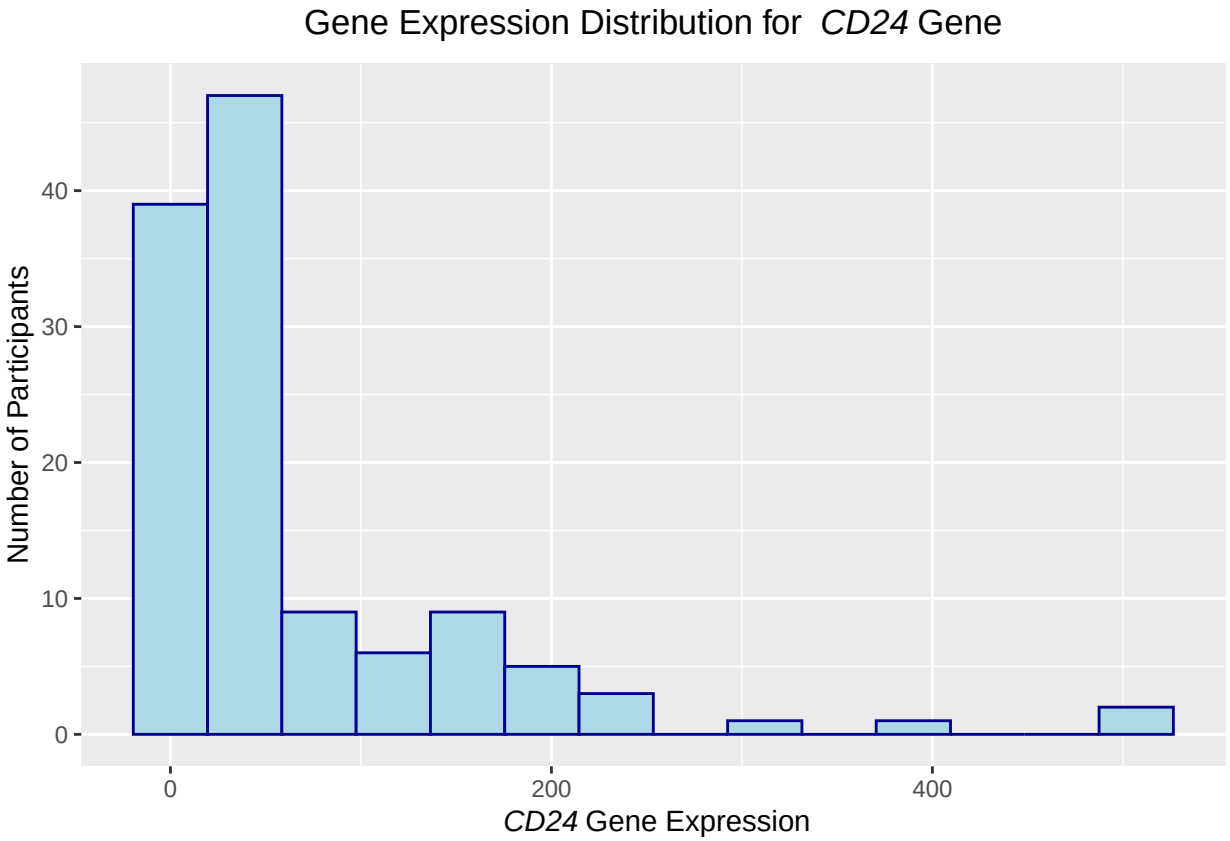

Gene Expression Distribution for LCN2 Gene
Number of Participants
30
20
10
0
0
1000
2000
3000
LCN2 Gene Expression




```
## [1] "LCN2"
```

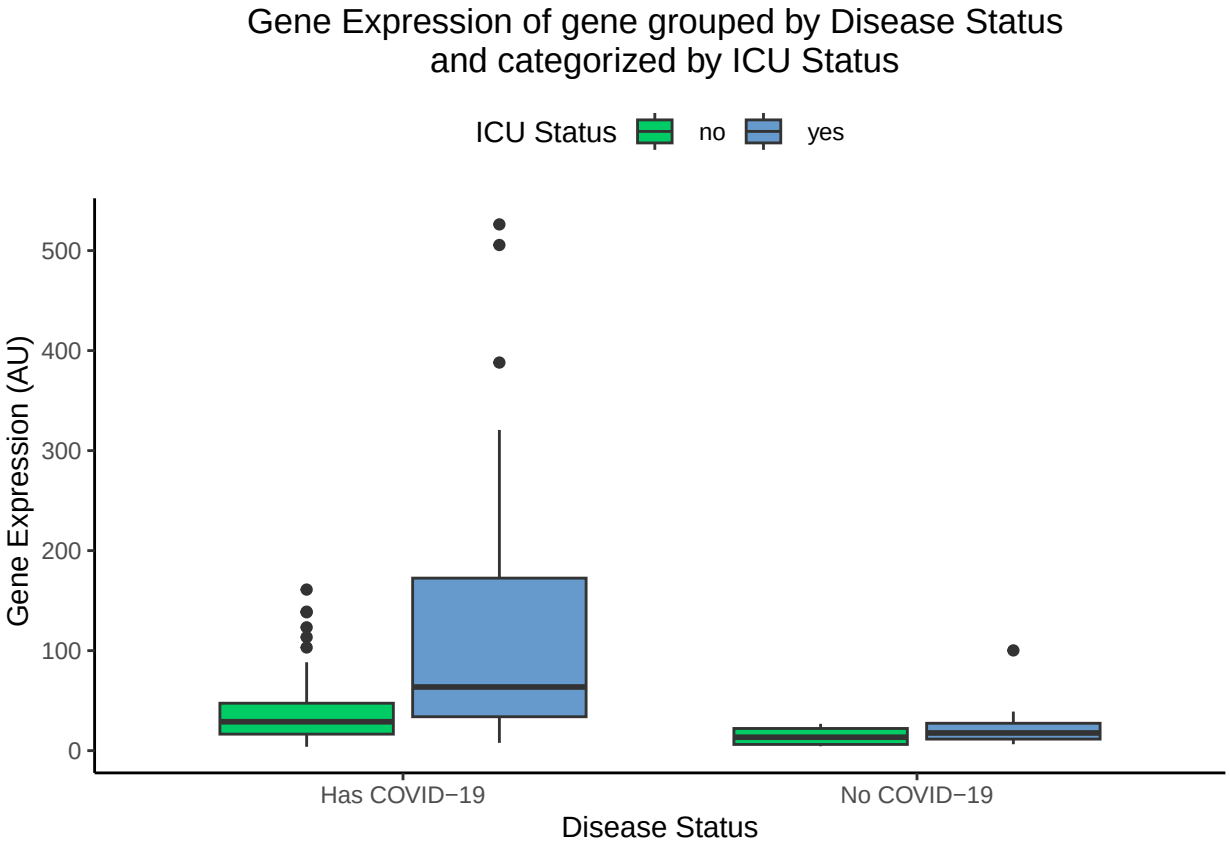
```
## Scale for colour is already present.
## Adding another scale for colour, which will replace the existing scale.
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```

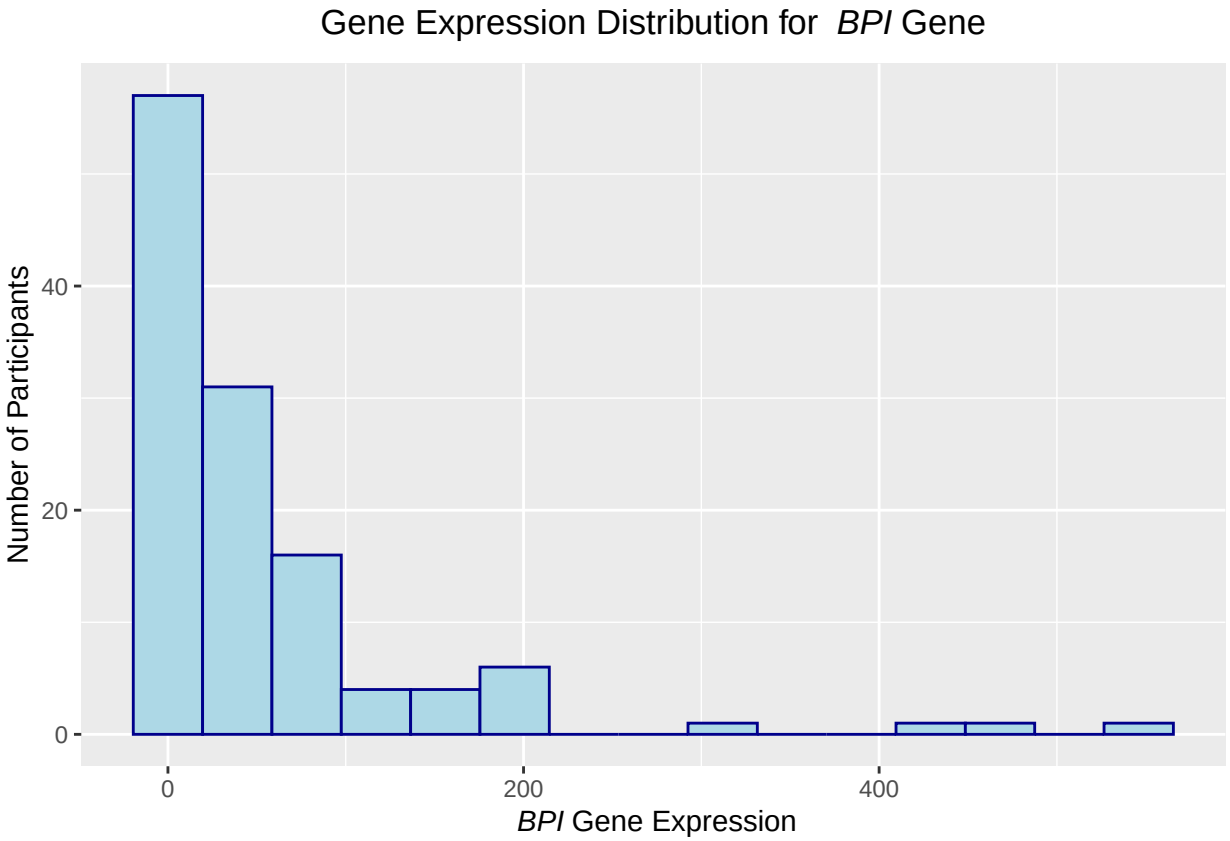

Gene Expression of gene grouped by Disease Status
and categorized by ICU Status
ICU Status
no
yes
3000
2000
1000
0
Gene Expression (AU)
Has COVID−19
No COVID−19
Disease Status

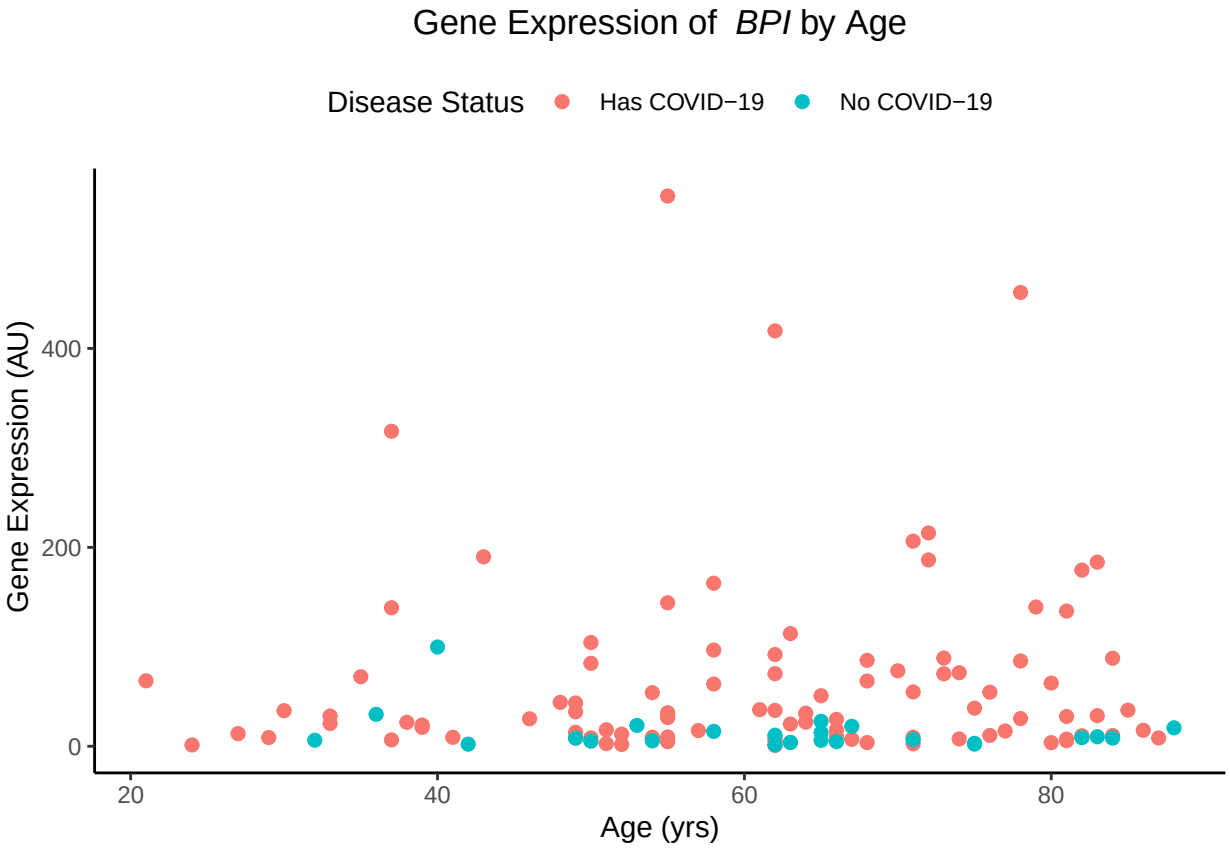

Gene Expression Distribution for CD24 Gene
Number of Participants
40
30
20
10
0
0
200
400
CD24 Gene Expression

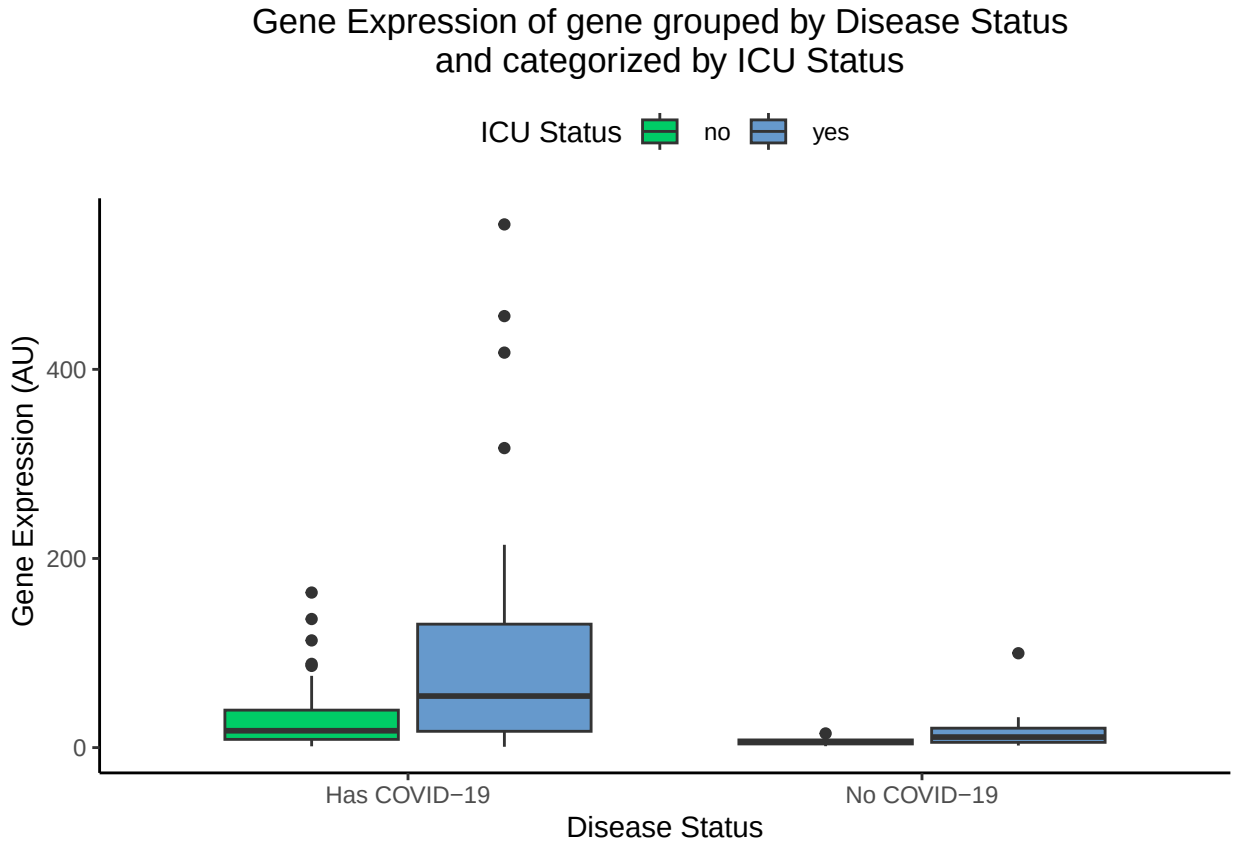

```
## [1] "CD24"
```

```
## Scale for colour is already present.
## Adding another scale for colour, which will replace the existing scale.
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```


Gene Expression of gene grouped by Disease Status
and categorized by ICU Status
ICU Status
no
yes
Gene Expression (AU)
500
400
300
200
100
0
Has COVID−19
No COVID−19
Disease Status


Gene Expression Distribution for BPI Gene
Number of Participants
40
20
0
0
200
400
BPI Gene Expression


Gene Expression of BPI by Age
Disease Status
Has COVID−19
No COVID−19
Gene Expression (AU)
0
200
400
20
40
60
80
Age (yrs)


Gene Expression of gene grouped by Disease Status
and categorized by ICU Status
ICU Status
no
yes
Gene Expression (AU)
400
200
0
Has COVID−19
No COVID−19
Disease Status

```
## [1] "BPI"
```