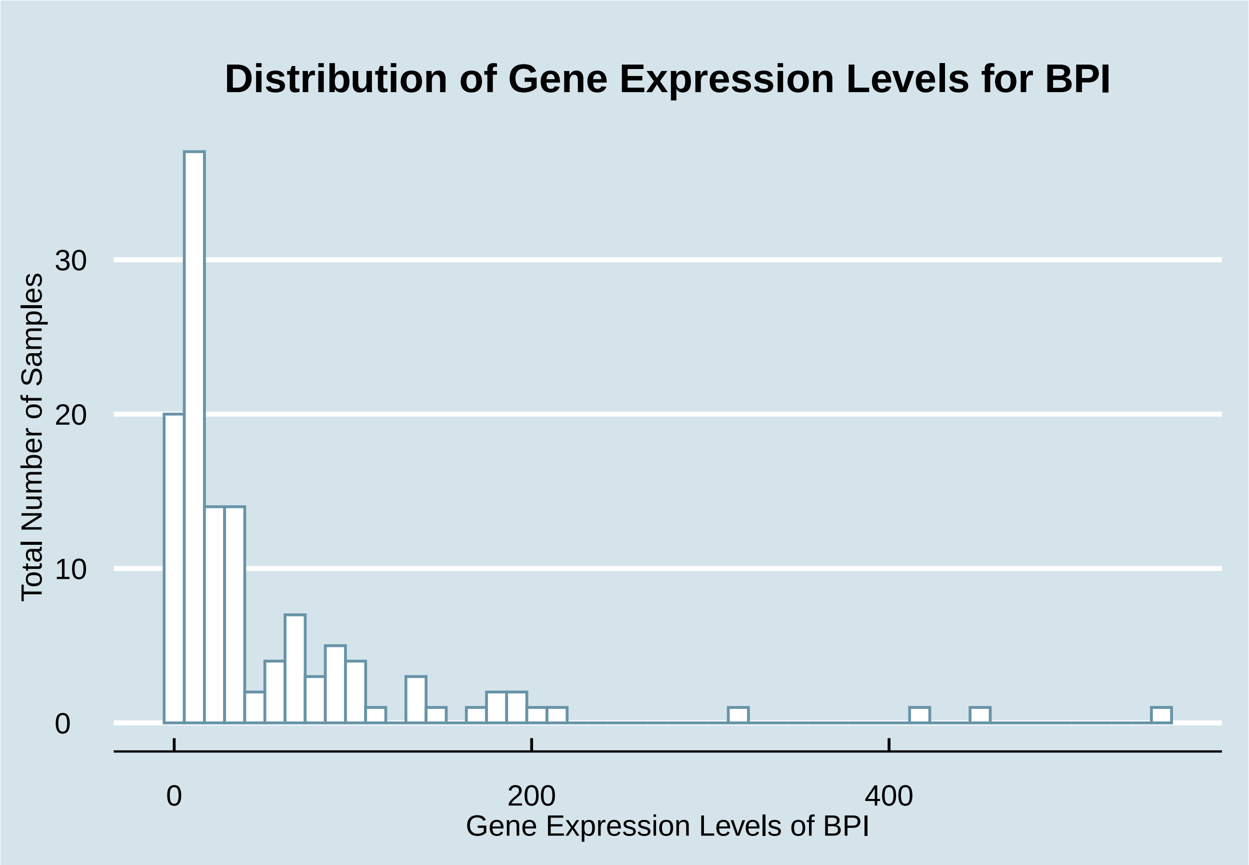
QBS103 Final Project

2023-07-23

```r
metadata <- read.csv("QBS103_finalProject_metadata.csv", row.names = 1)
gene_exp <- read.csv("QBS103_finalProject_geneExpression.csv", row.names = 1)
```

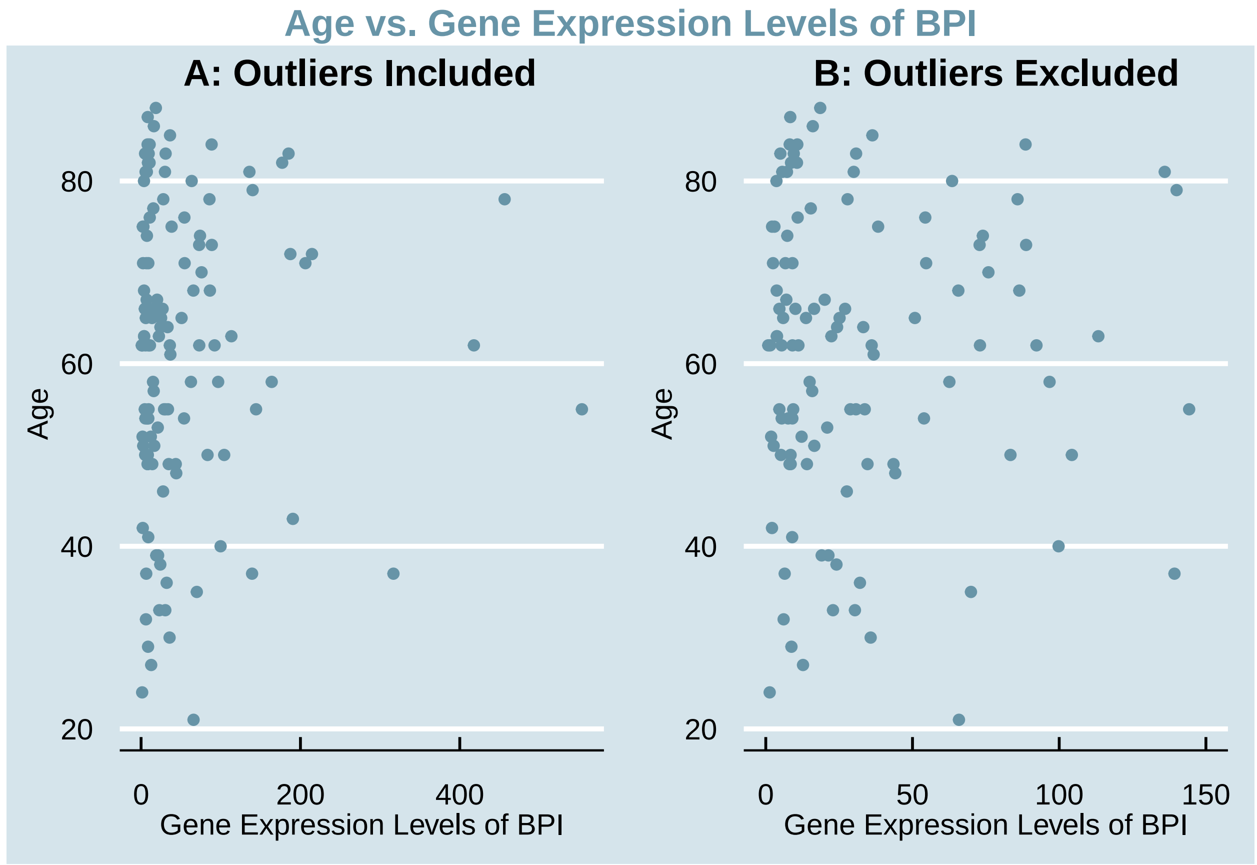
```r
#chosen gene: BPI - bactericidal permeability increasing protein - Plays a role in the immune response

BPI <- gene_exp["BPI",]
BPI <- as.data.frame(t(BPI)) #transposes the matrix
metadata_BPI <- cbind(metadata, BPI) #creates one data frame with all BPI data
suppressWarnings(metadata_BPI$age <- as.integer(metadata_BPI$age))
```

```r
#creates histogram of BPI

ggplot(metadata_BPI, aes(x = BPI)) + geom_histogram(color = "#6794a7", fill = "white", bins = 50) +
  labs(x = "Gene Expression Levels of BPI", y = "Total Number of Samples", title = "Distribution of Gen
  theme_economist() +
  scale_fill_economist() +
  theme(plot.title = element_text(size=15, face="bold", margin = margin(10, 0, 10, 0), hjust = (0.5)),
```



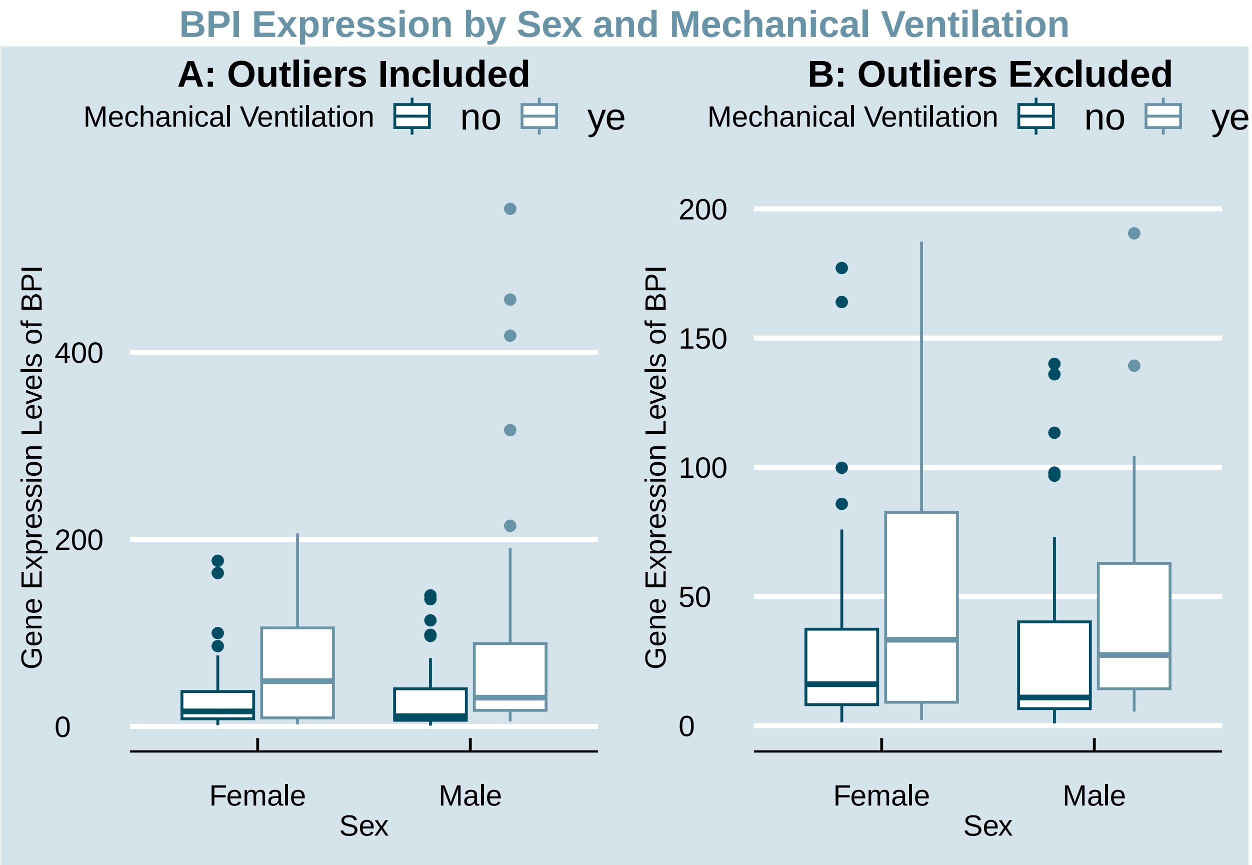
```
# interpretation - shows the distribution of BPI gene expression in the samples. there are more studies
```

```
#chosen continuous covariate: age

#creates scatterplot with BPI and age
suppressWarnings(scatter_outliers <- ggplot(metadata_BPI, aes(x = BPI, y = age)) +
  geom_point(color = "#6794a7", fill = "white") +
  labs(x = "Gene Expression Levels of BPI", y = "Age") +
  theme_economist() +
  scale_fill_economist() +
  theme(plot.title = element_text(size=12, face="bold", margin = margin(10, 0, 10, 0)), axis.title.x = e

#creates scatterplot with BPI and age and sets x-axis range from 0 to 150 for easier viewing
suppressWarnings(scatter_no_outliers <- ggplot(metadata_BPI, aes(x = BPI, y = age)) +
  geom_point(color = "#6794a7", fill = "white") +
  labs(x = "Gene Expression Levels of BPI", y = "Age") +
  theme_economist() +
  scale_fill_economist() +
  theme(plot.title = element_text(size=12, face="bold", margin = margin(10, 0, 10, 0)), axis.title.x = e
  xlim(0, 150))

#creates scatterplot of both previous plots combined
suppressWarnings(new_scatter <- ggarrange(scatter_outliers, scatter_no_outliers,ncol=2, labels = c("A: (

#adds title to the combined plot
suppressWarnings(annotate_figure(new_scatter, top = text_grob("Age vs. Gene Expression Levels of BPI", c
```



```
#interpretation - shows the relationship between BPI gene expression and ages of individuals sampled. t
```

```
#chosen categorical covariates: sex and mechanical ventilation

#create new dataset with the row of data where sex = "unknown"
unknown_removed <- metadata_BPI
unknown_removed <- metadata_BPI[metadata_BPI$sex != " unknown", ]

#creates scatterplot with BPI, sex, and mechanical ventilation
suppressWarnings(box_outliers <- ggplot(unknown_removed,aes(x = sex,y = BPI, color = mechanical_ventila

#creates scatterplot with BPI, sex, and mechanical ventilation and sets y-axis range from 0 to 200 for
suppressWarnings(box_no_outliers <- ggplot(unknown_removed,aes(x = sex,y = BPI, color = mechanical_venti

#creates boxplot of both previous plots combined
suppressWarnings(new_box <- ggarrange(box_outliers, box_no_outliers,ncol=2, labels = c("A: Outliers Incl

#adds title to the combined plot
annotate_figure(new_box, top = text_grob("BPI Expression by Sex and Mechanical Ventilation", color = "#
```



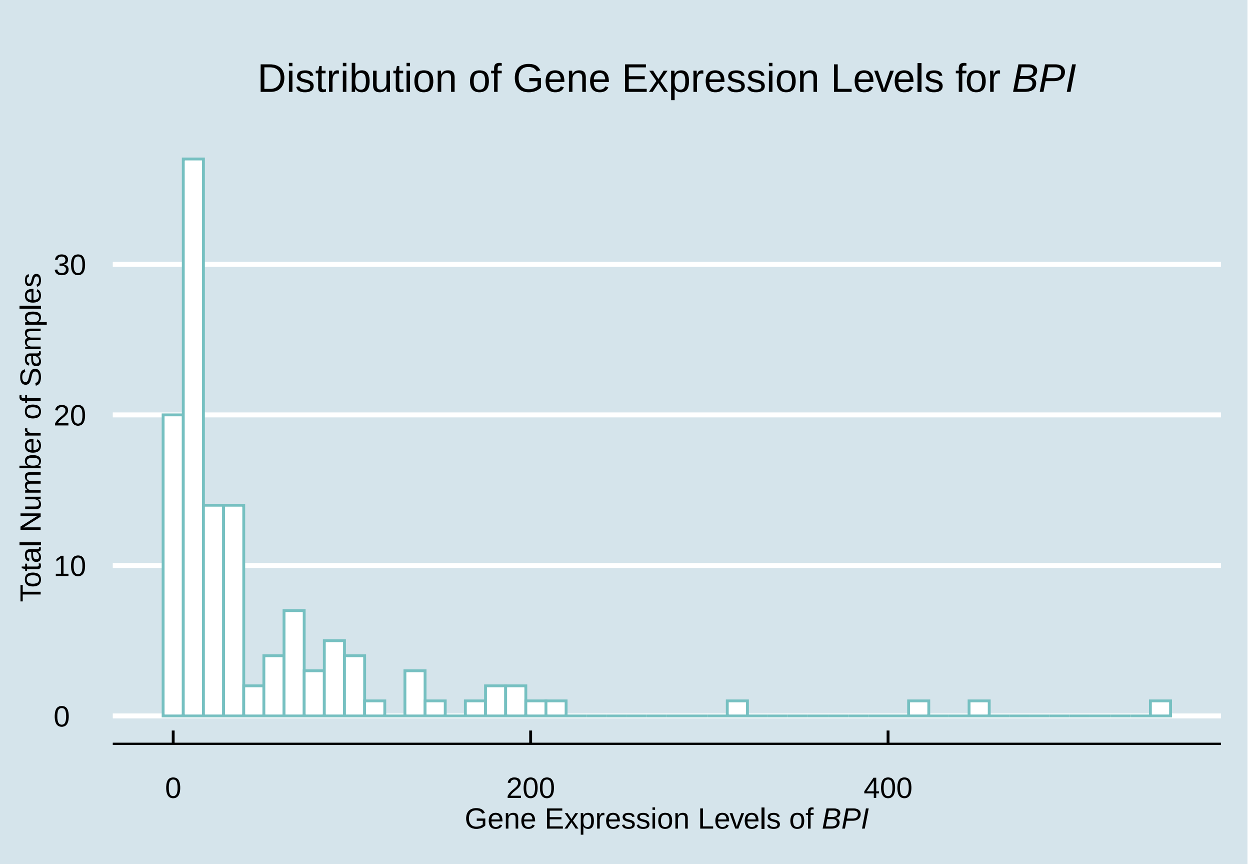
```
#interpretation - shows distribution of BPI gene expression across sexes and ventilation categories. fr
```

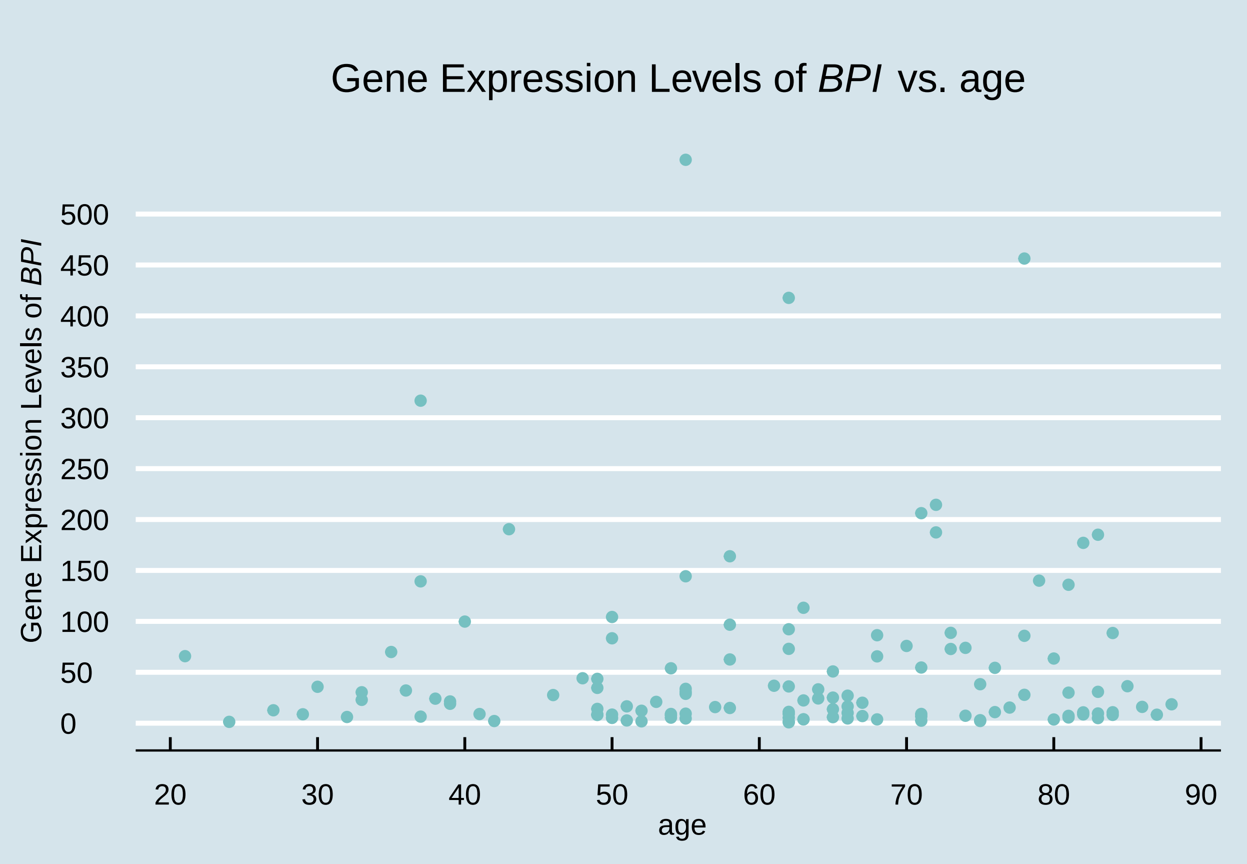
```
#Code for boxplot, scatter, and histogram
all_graphs <- function(data, genes, gene_exp, contcovariate, catcovariate1, catcovariate2, metadata) {
  graphs <- list()
  for (gene in genes) { #creates a loop to create each graph per gene
    gene_data <- as.numeric(t(gene_exp[gene, ]))
    metadata_gene <- cbind(metadata, gene_data)

    #relabels data within a few covariates for better keys in boxplots
    metadata_gene$icu_status <- ifelse(metadata_gene$icu_status == " yes", "in icu", "not in icu")
    metadata_gene$mechanical_ventilation <- ifelse(metadata_gene$mechanical_ventilation == " yes", "on r
    metadata_gene$disease_status <- ifelse(metadata_gene$disease_status == "disease state: COVID-19", "


    #creates histogram
    hist <- ggplot(metadata_gene, aes(x = gene_data)) +
      geom_histogram(color = "#76c0c1", fill = "white", bins = 50) +
      labs(x = substitute(paste("Gene Expression Levels of ", italic(gene)), list(gene = gene)), #subst
           y = "Total Number of Samples",
           title = substitute(paste("Distribution of Gene Expression Levels for ", italic(gene)), list(
      theme_economist() +
      scale_fill_economist() +
      theme(plot.title = element_text(size = 15, face = "bold", margin = margin(10, 0, 10, 0), hjust =
            axis.title.x = element_text(vjust = -0.4),
            axis.title.y = element_text(vjust = 2))
```

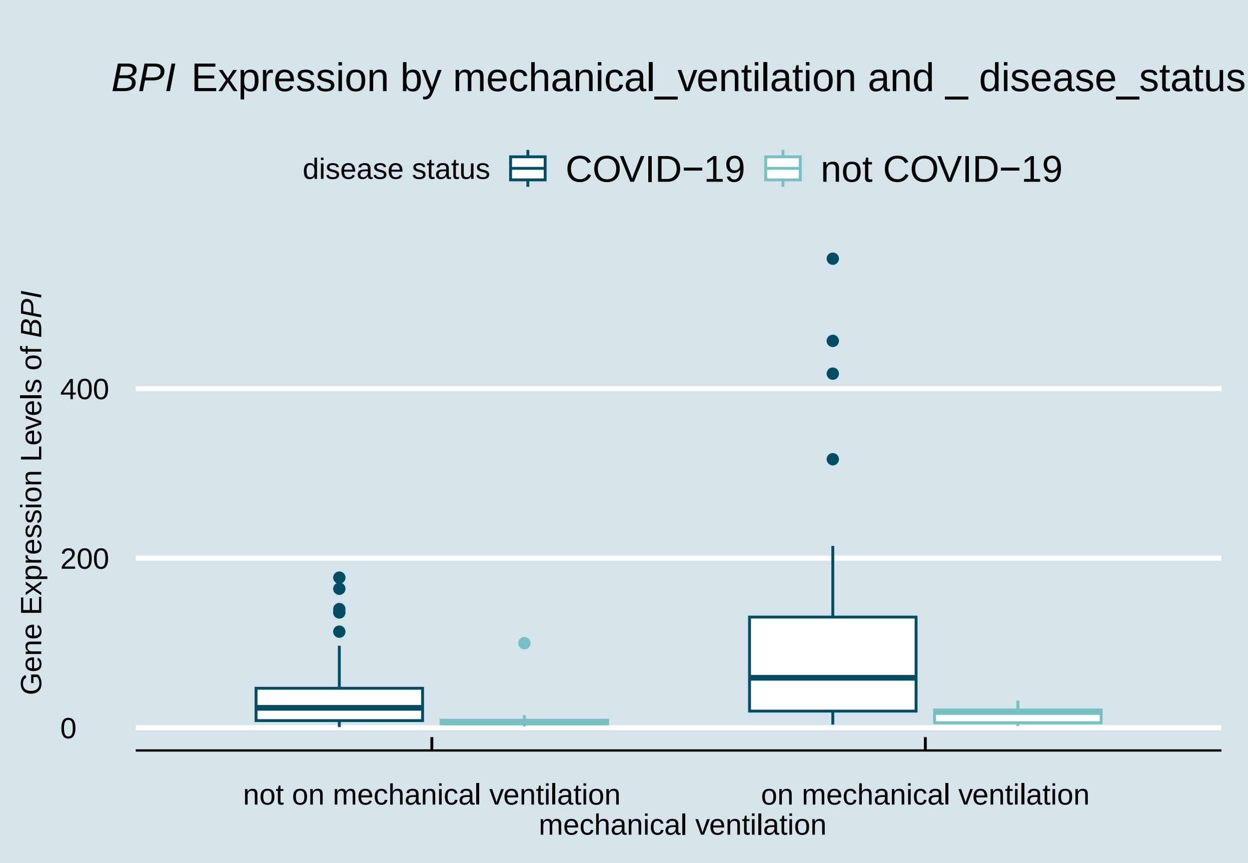
```r
    #creates scatter plot
    metadata_gene$contcovariate <- as.integer(metadata_gene[[contcovariate]])
    metadata_gene <- metadata_gene[!is.na(metadata_gene$contcovariate), ]

    scatter <- ggplot(metadata_gene, aes(x = contcovariate, y = gene_data)) +
      geom_point(color = "#76c0c1", fill = "white") +
      labs(x = paste("", contcovariate) , y = substitute(paste("Gene Expression Levels of ", italic(gen
      theme_economist() +
      scale_fill_economist() +
      theme(plot.title = element_text(size = 15, face = "bold", margin = margin(10, 0, 10, 0), hjust =
      ggtitle(substitute(paste("Gene Expression Levels of ", italic(gene), " vs. ", contcovariate), lis


    #creates boxplot
    metadata_gene$catcovariate1 <- as.factor(metadata_gene[[catcovariate1]])
    metadata_gene$catcovariate2 <- as.factor(metadata_gene[[catcovariate2]])
    metadata_gene <- metadata_gene[!is.na(metadata_gene$catcovariate1) & !is.na(metadata_gene$catcovaria

    box <- ggplot(metadata_gene, aes(x = catcovariate1, y = gene_data, color = catcovariate2)) +
      geom_boxplot() +
      scale_x_discrete(labels = levels(metadata_gene$catcovariate1)) +
      labs(x = gsub("_", " ", paste("", catcovariate1)), y = substitute(paste("Gene Expression Levels o
      theme_economist() +
      scale_color_manual(values = c('#014d64','#76c0c1')) +
      theme(plot.title = element_text(size = 15, face ="bold", margin = margin(10, 0, 10, 0), hjust = (
      ggtitle(substitute(paste(italic(gene), " Expression by ", catcovariate1, " and ", "_", " ", catco

    plots <- list(hist = hist, scatter = scatter, box = box)
    graphs[[gene]] <- plots

  }
  return(graphs)
}

selected_genes <- c("BPI", "MPO")

multiple_graphs <- all_graphs(data = gene_exp, genes = selected_genes, contcovariate = "age", catcovaria
```

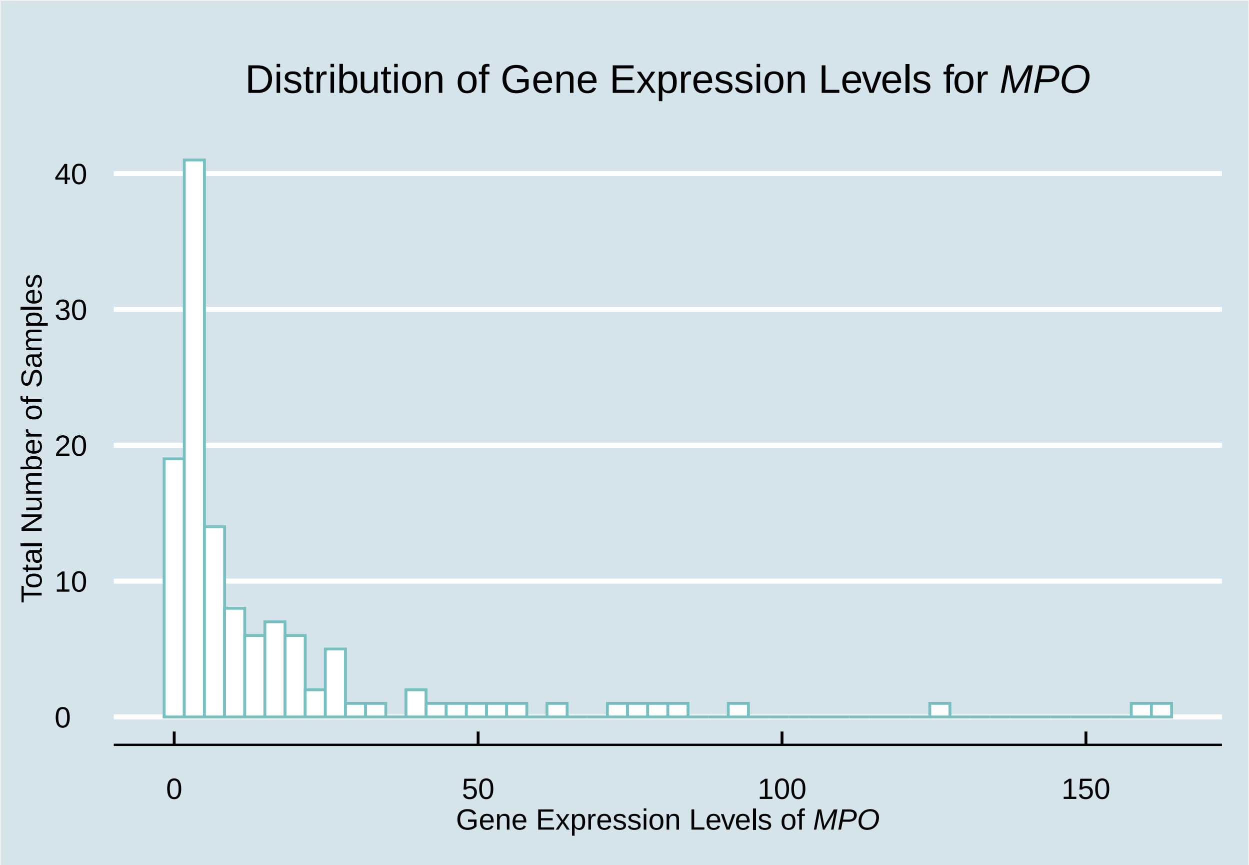
```
## Warning in all_graphs(data = gene_exp, genes = selected_genes, contcovariate =
## "age", : NAs introduced by coercion

## Warning in all_graphs(data = gene_exp, genes = selected_genes, contcovariate =
## "age", : NAs introduced by coercion
```

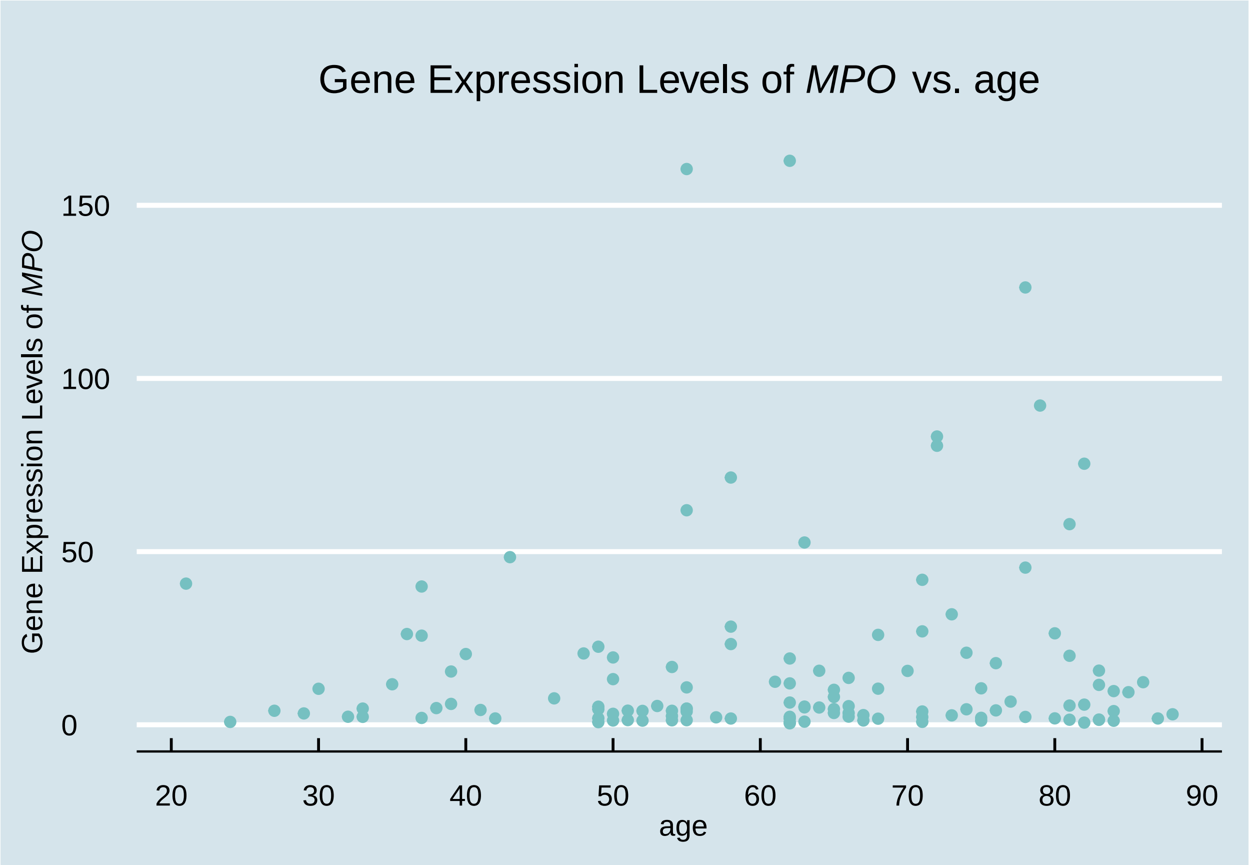
```r
for (gene in selected_genes) {
  print(multiple_graphs[[gene]]$hist)
  print(multiple_graphs[[gene]]$scatter)
  print(multiple_graphs[[gene]]$box)
}
```

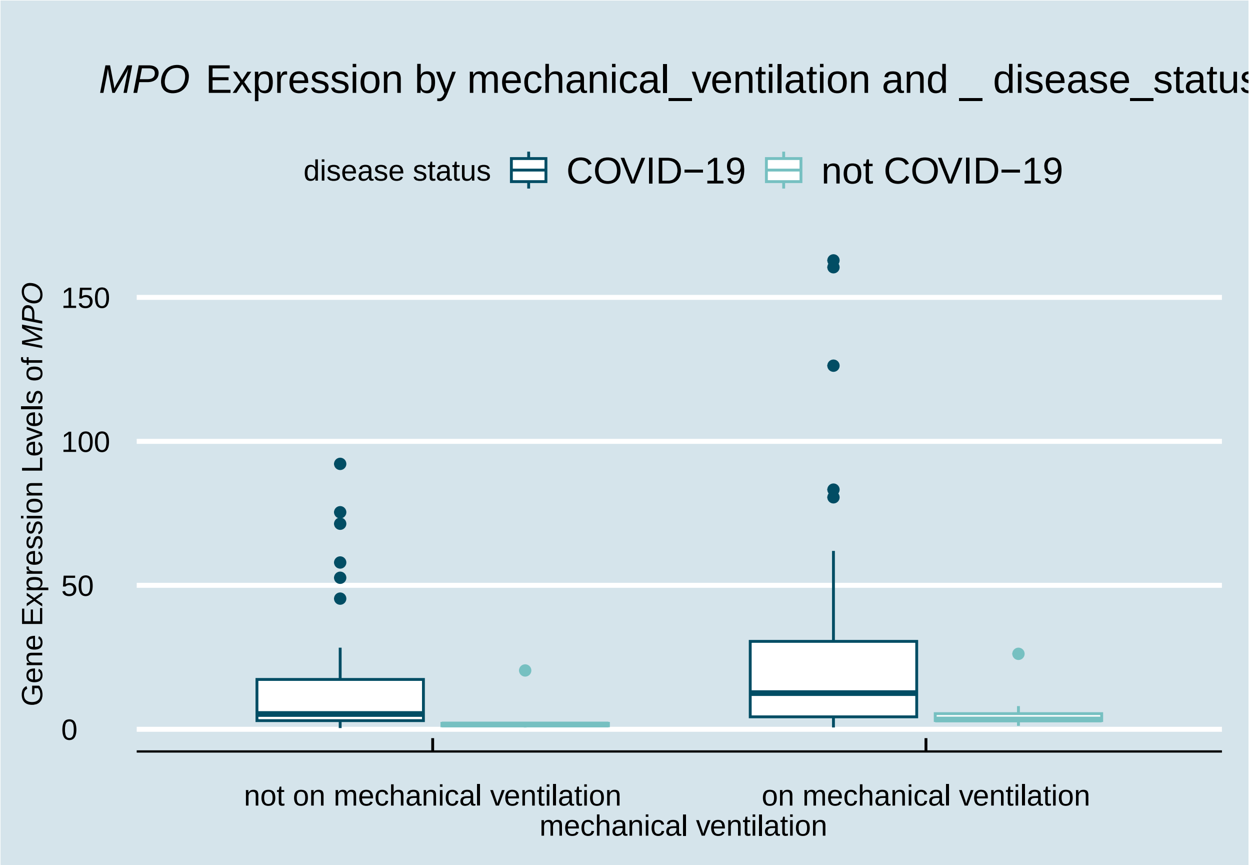

Distribution of Gene Expression Levels for *BPI*
Total Number of Samples
0
10
20
30
0
200
400
Gene Expression Levels of *BPI*


Gene Expression Levels of *BPI* vs. age
Gene Expression Levels of *BPI*
500
450
400
350
300
250
200
150
100
50
0
20
30
40
50
60
70
80
90
age


BPI Expression by mechanical_ventilation and _ disease_status
disease status
COVID−19
not COVID−19
Gene Expression Levels of BPI
400
200
0
not on mechanical ventilation
on mechanical ventilation
mechanical ventilation


Distribution of Gene Expression Levels for *MPO*
Total Number of Samples
0
10
20
30
40
0
50
100
150
Gene Expression Levels of *MPO*


Gene Expression Levels of *MPO* vs. age
Gene Expression Levels of *MPO*
150
100
50
0
20
30
40
50
60
70
80
90
age

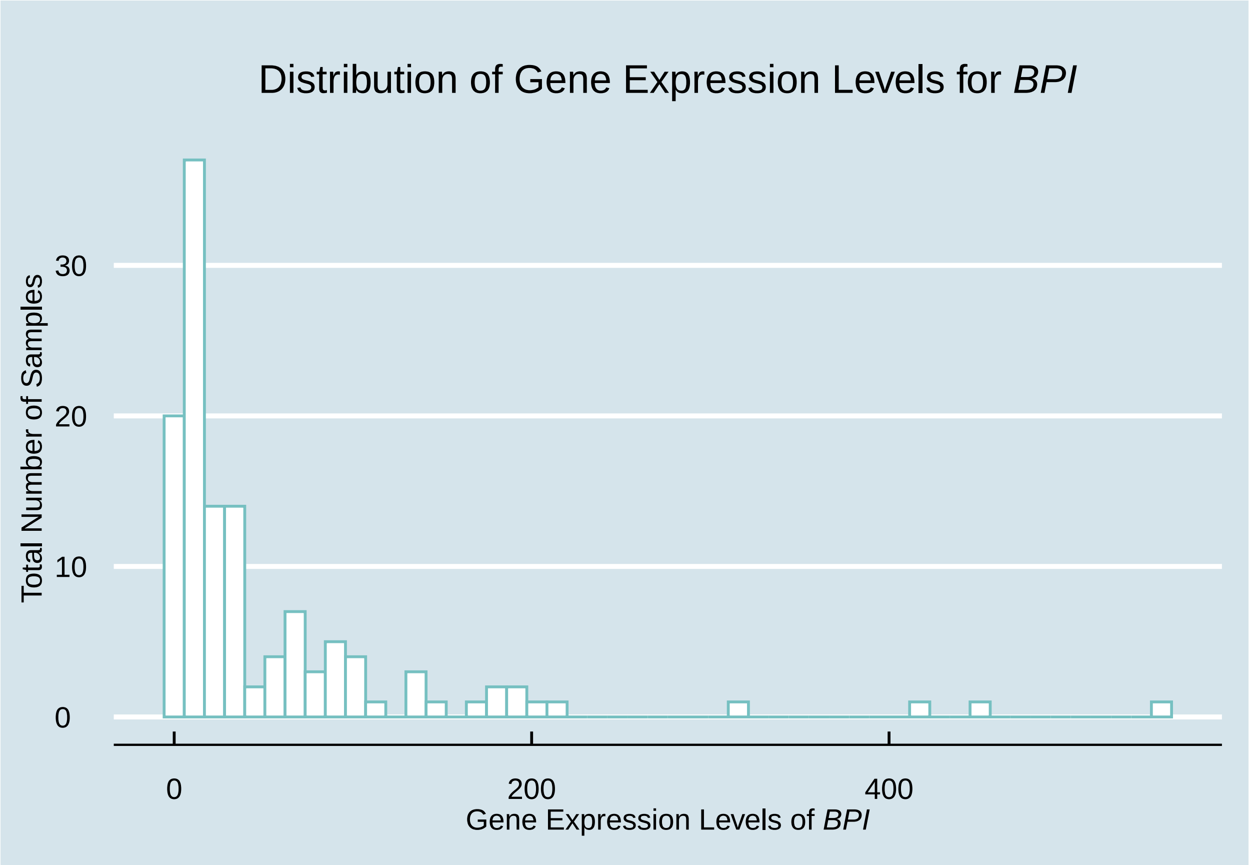


```
#calling the function
additional_genes <- c("GPLD1", "AAK1")
selected_genes <- c("BPI", "MPO", "GPLD1", "AAK1")
multiple_graphs <- all_graphs(data = gene_exp, genes = selected_genes, contcovariate = "age", catcovari
```

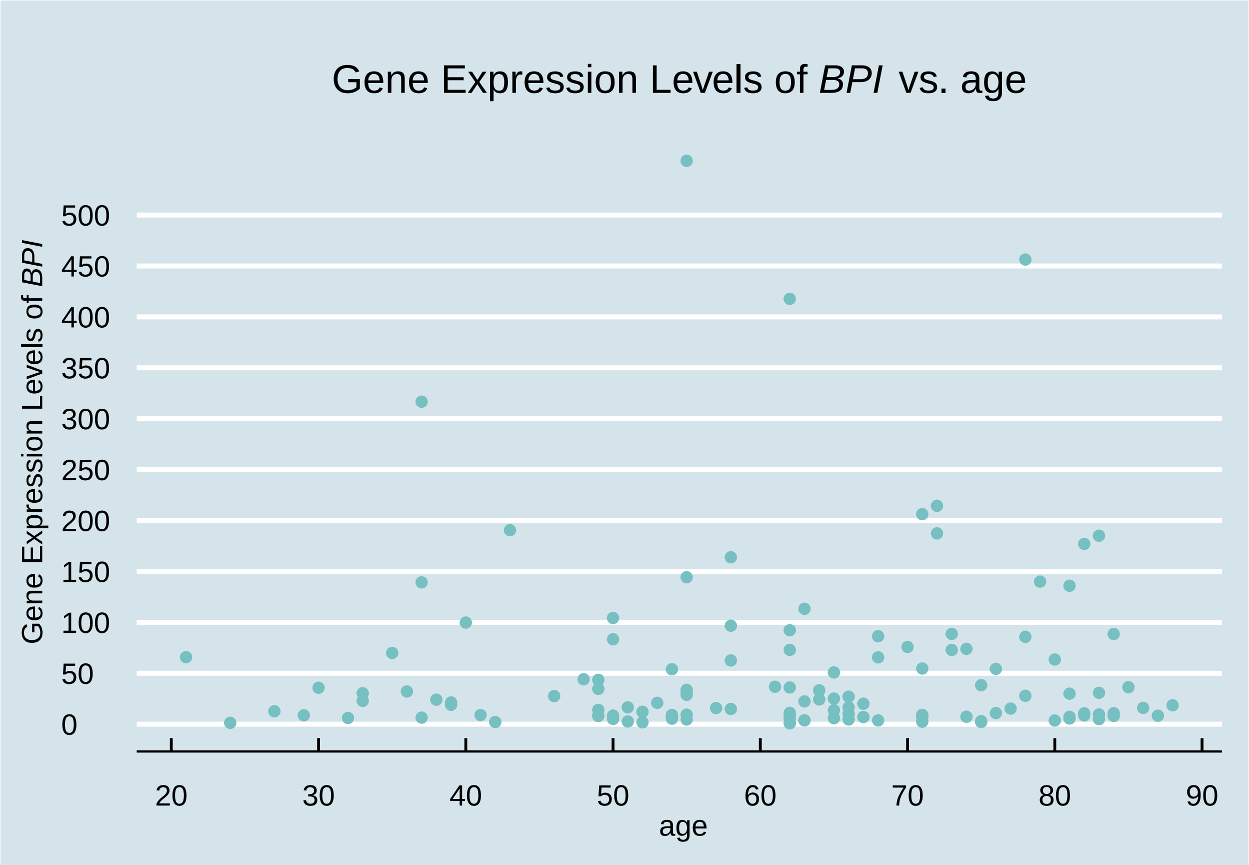
```
## Warning in all_graphs(data = gene_exp, genes = selected_genes, contcovariate =
## "age", : NAs introduced by coercion

## Warning in all_graphs(data = gene_exp, genes = selected_genes, contcovariate =
## "age", : NAs introduced by coercion

## Warning in all_graphs(data = gene_exp, genes = selected_genes, contcovariate =
## "age", : NAs introduced by coercion

## Warning in all_graphs(data = gene_exp, genes = selected_genes, contcovariate =
## "age", : NAs introduced by coercion
```

```
print(multiple_graphs)
```

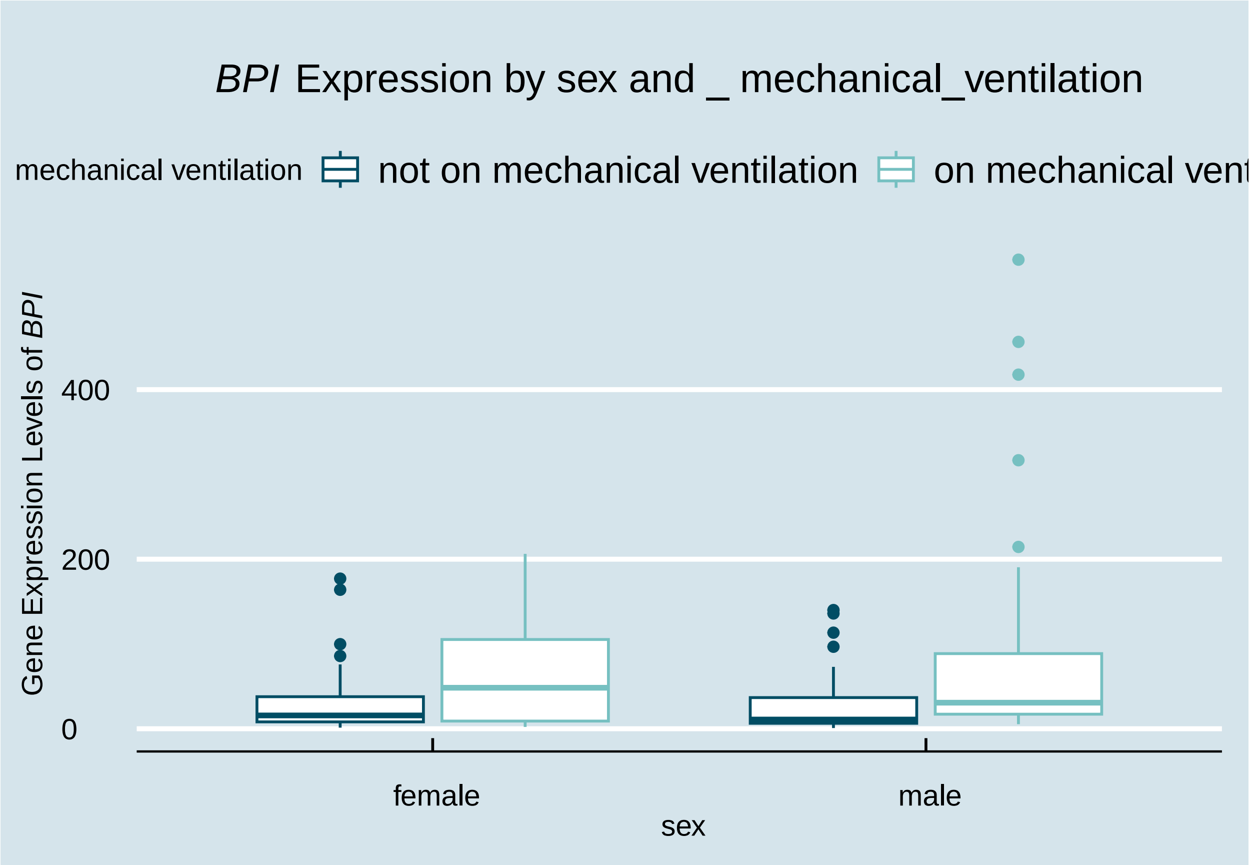
```
## $BPI
## $BPI$hist
```


Distribution of Gene Expression Levels for *BPI*
Total Number of Samples
0
10
20
30
0
200
400
Gene Expression Levels of *BPI*

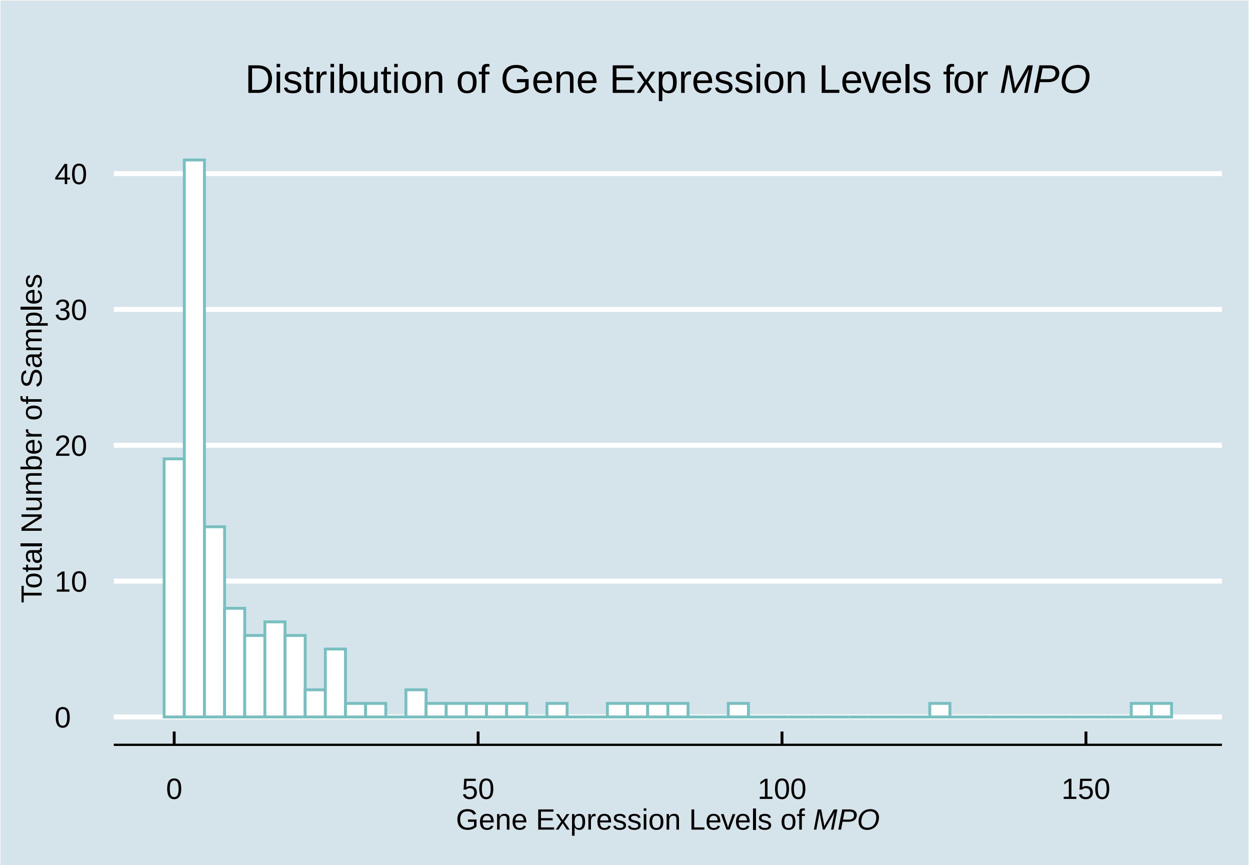
```
##
## $BPI$scatter
```


Gene Expression Levels of *BPI* vs. age
Gene Expression Levels of *BPI*
500
450
400
350
300
250
200
150
100
50
0
20
30
40
50
60
70
80
90
age

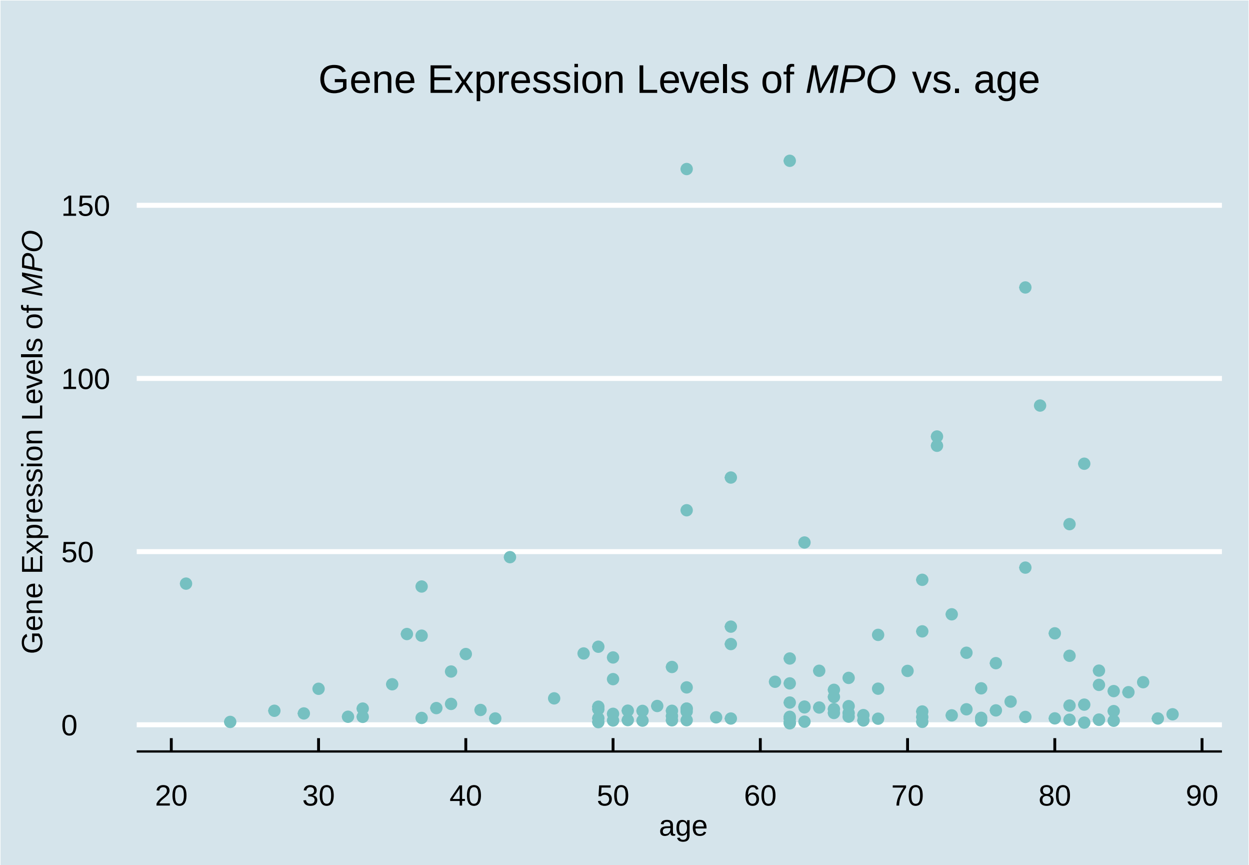
```
##
## $BPI$box
```


BPI Expression by sex and _ mechanical_ventilation
mechanical ventilation
not on mechanical ventilation
on mechanical vent
Gene Expression Levels of *BPI*
0
200
400
female
male
sex

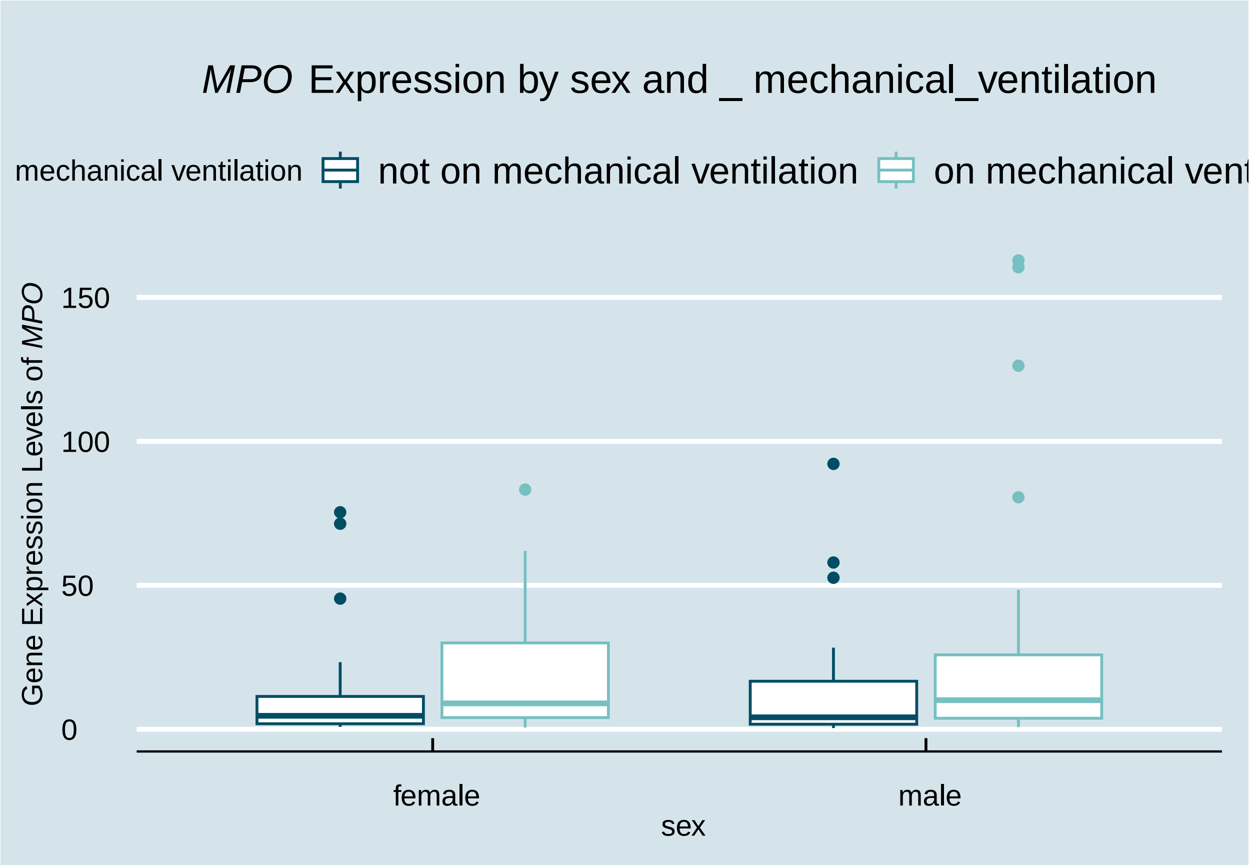
```
##
##
## $MPO
## $MPO$hist
```


Distribution of Gene Expression Levels for *MPO*
Total Number of Samples
0
10
20
30
40
0
50
100
150
Gene Expression Levels of *MPO*

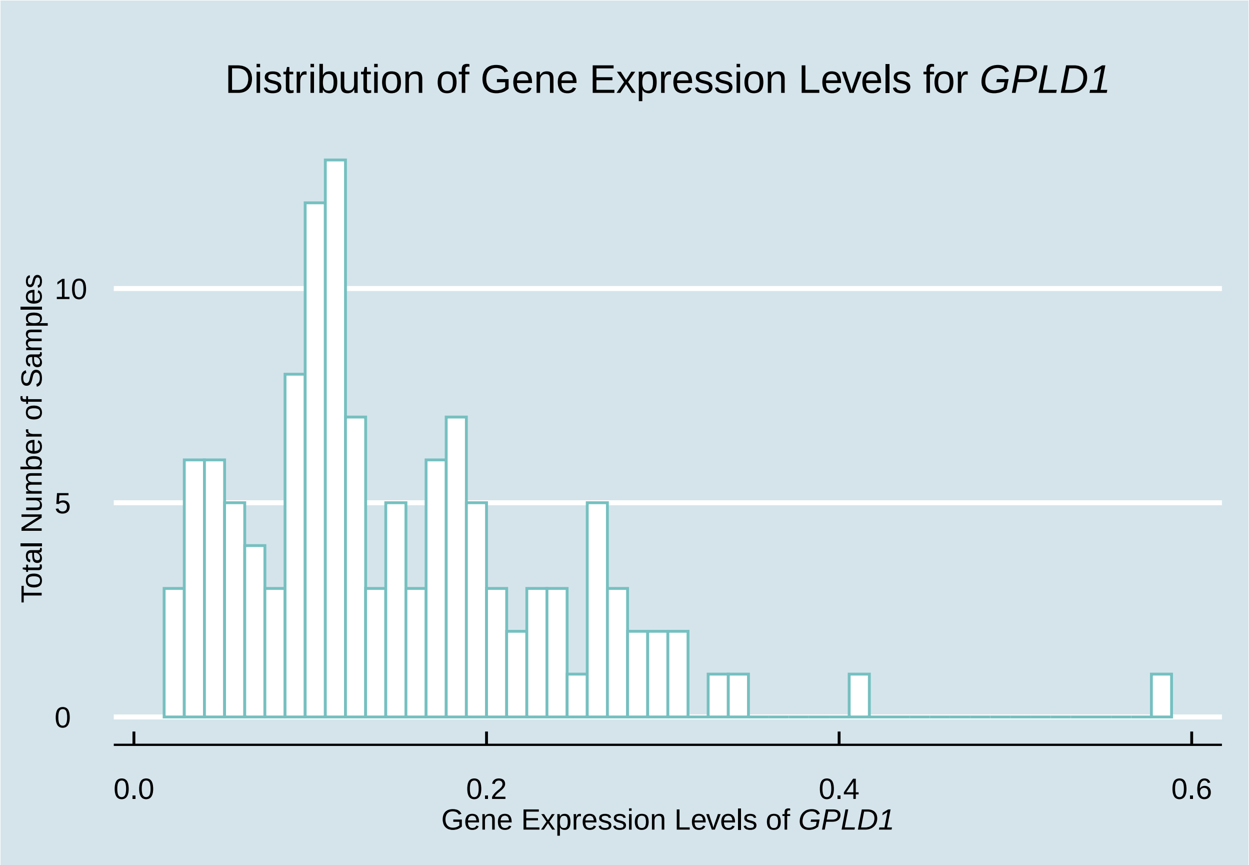
```
##
## $MPO$scatter
```


Gene Expression Levels of *MPO* vs. age
Gene Expression Levels of *MPO*
0
50
100
150
20
30
40
50
60
70
80
90
age

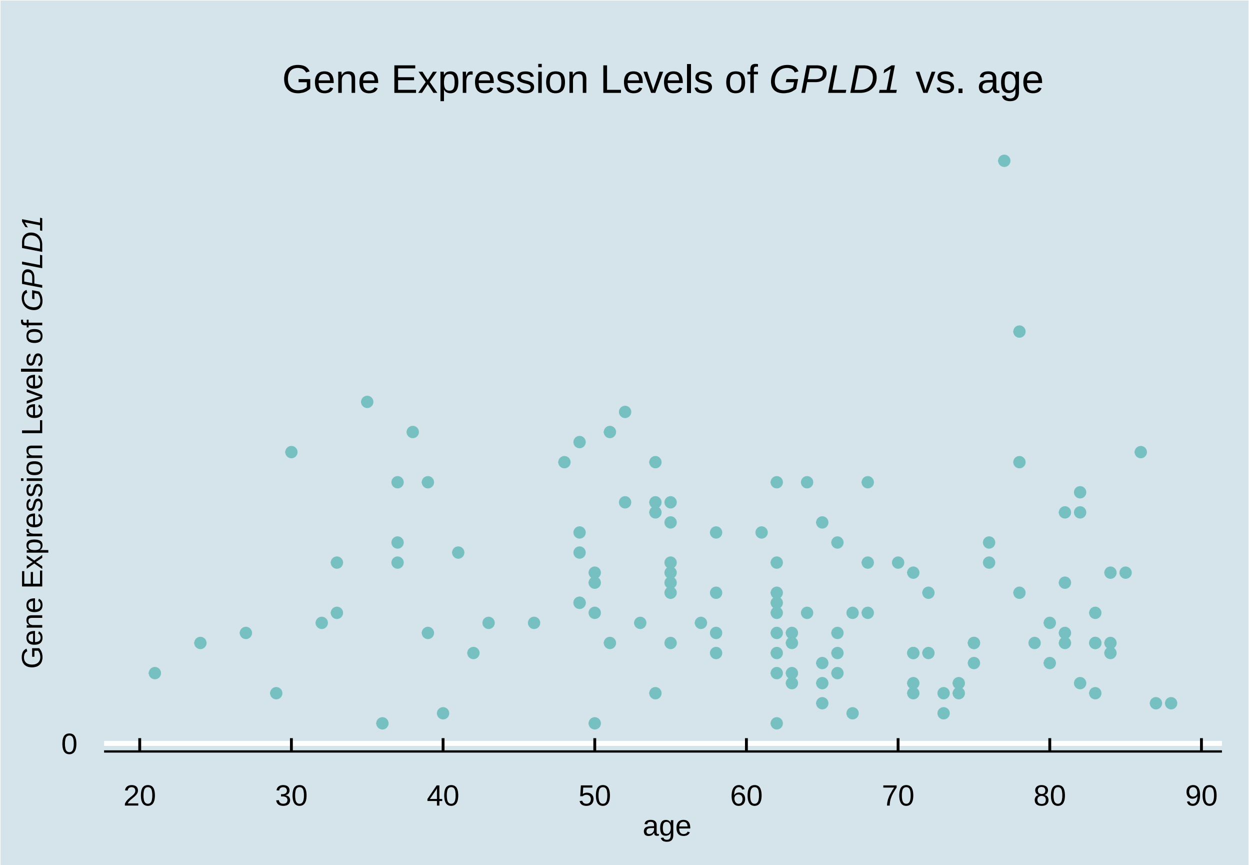
```
##
## $MPO$box
```



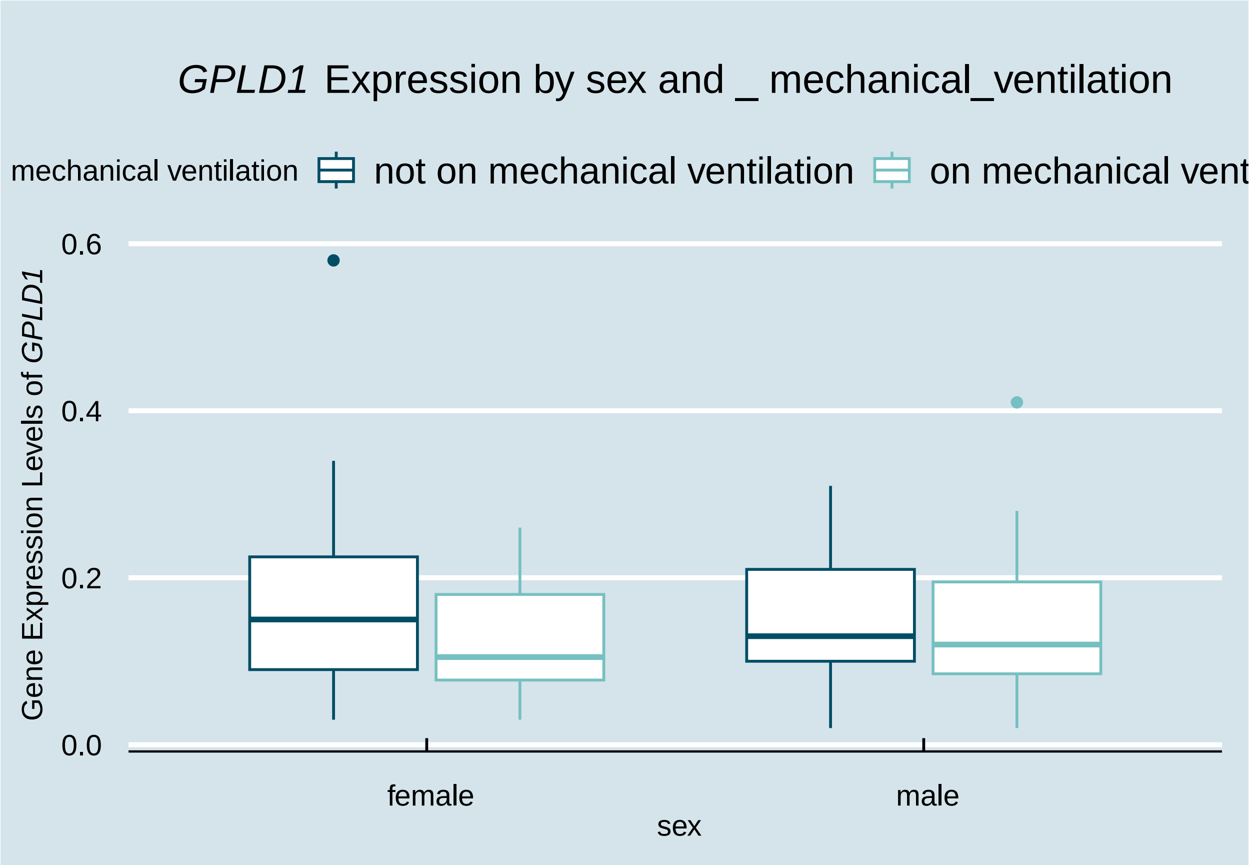
```
##
##
## $GPLD1
## $GPLD1$hist
```



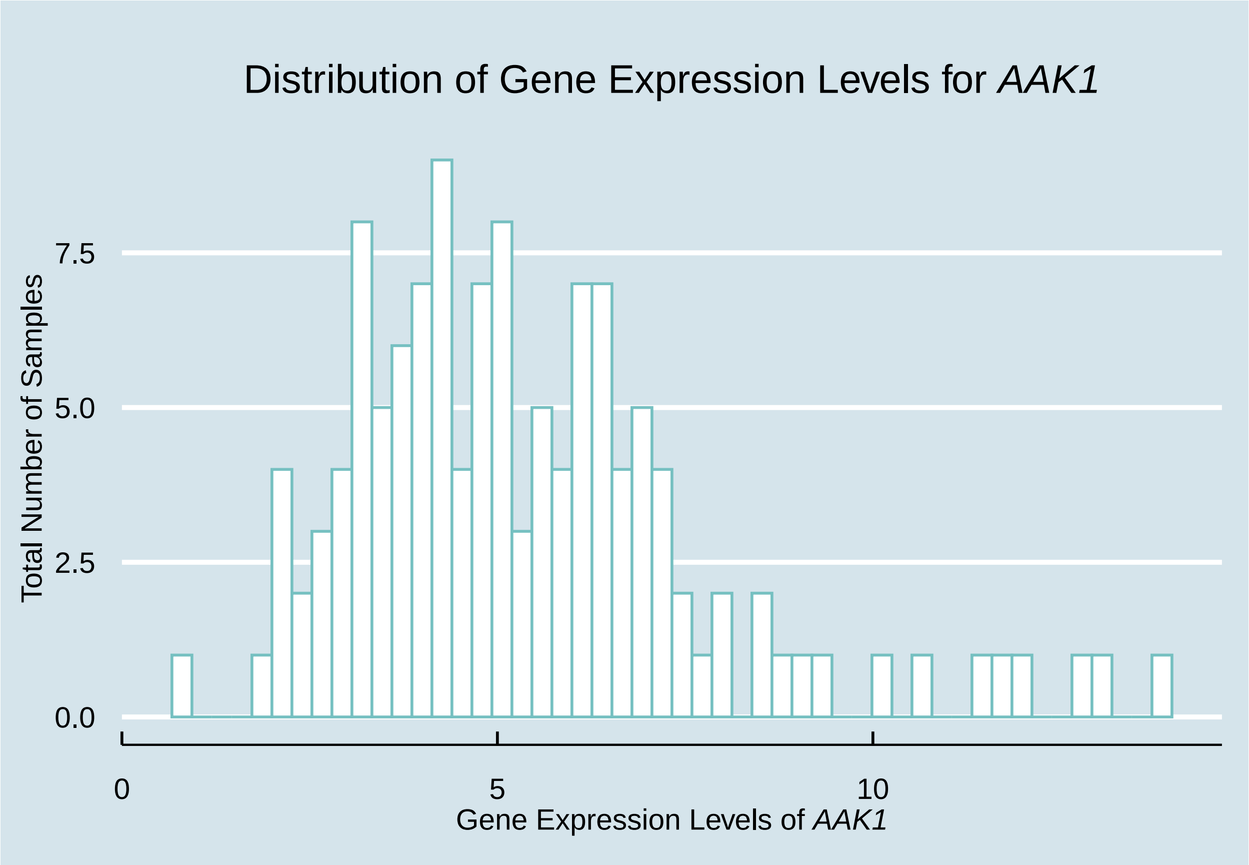
```
##
## $GPLD1$scatter
```


Gene Expression Levels of *GPLD1* vs. age
Gene Expression Levels of *GPLD1*
0
20
30
40
50
60
70
80
90
age

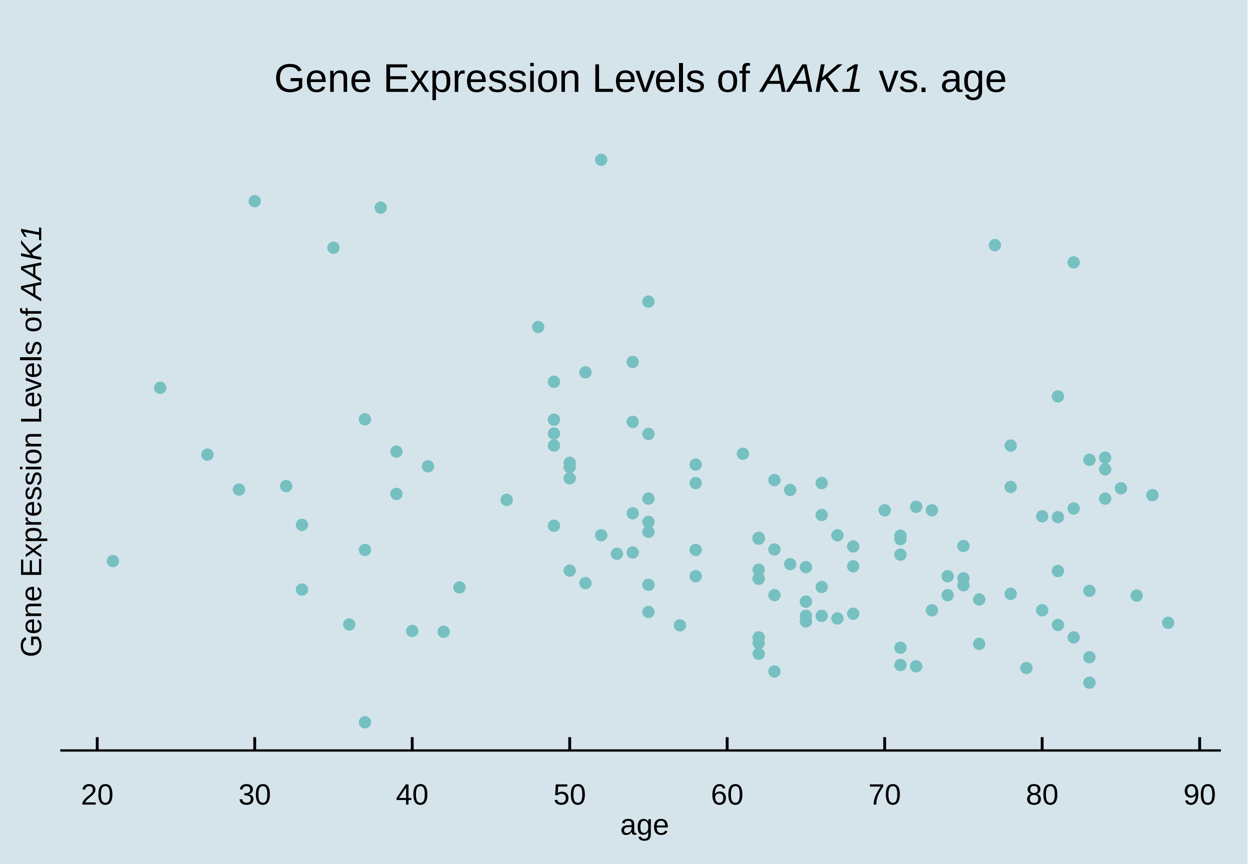
```
##
## $GPLD1$box
```



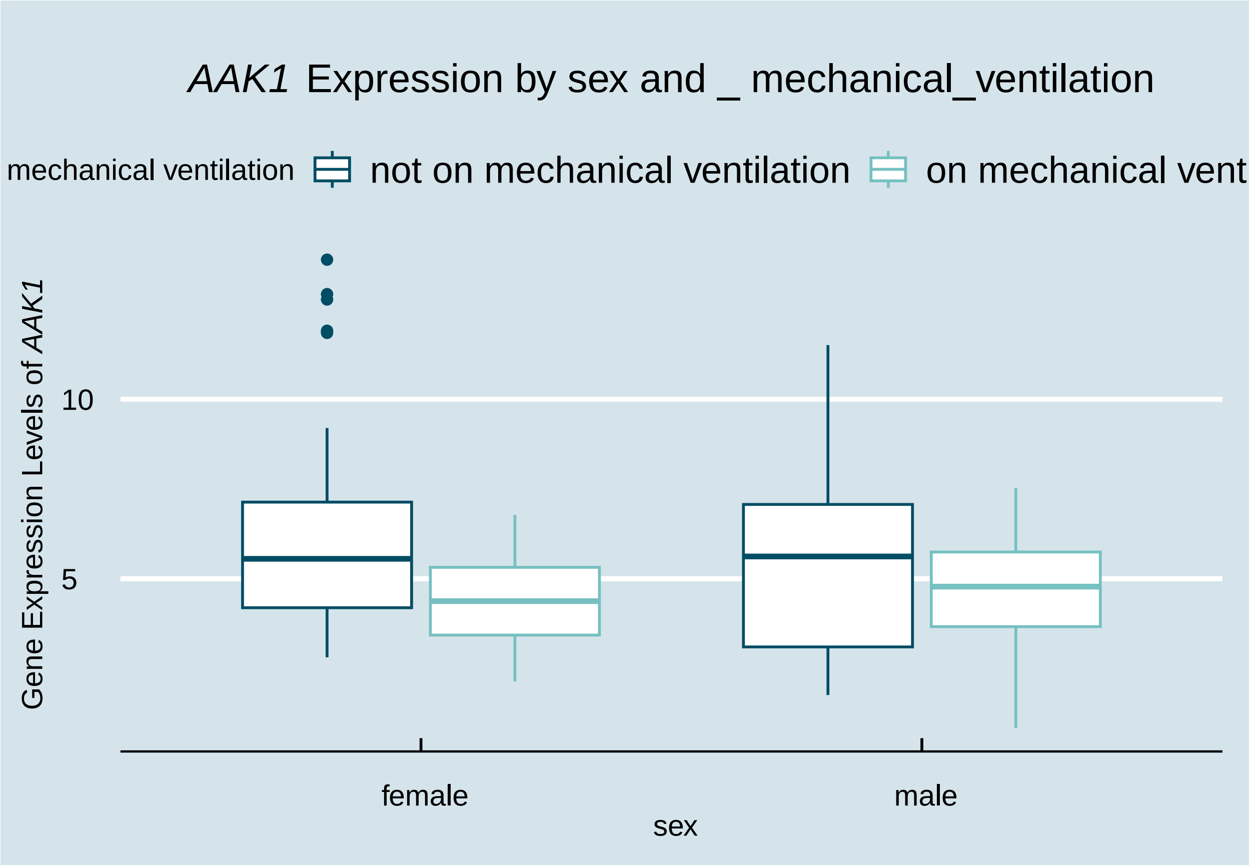
```
##
##
## $AAK1
## $AAK1$hist
```



```
##
## $AAK1$scatter
```


Gene Expression Levels of *AAK1* vs. age
Gene Expression Levels of *AAK1*
20
30
40
50
60
70
80
90
age

```
##
## $AAK1$box
```


AAK1 Expression by sex and _ mechanical_ventilation
mechanical ventilation
not on mechanical ventilation
on mechanical vent
Gene Expression Levels of *AAK1*
10
5
female
male
sex