QBS103 Project - Elodie Richard

2024-07-25

```
project1.data <- list.files(path = "/Users/elodierichard/Documents/QBS103/Project Submission 1 Data", pa
print(project1.data) #I first moved both data files into one folder on my laptop to retrieve it
```

```
## [1] "QBS103_GSE157103_genes.csv"         "QBS103_GSE157103_series_matrix.csv"
```

```
setwd("/Users/elodierichard/Documents/QBS103/Project Submission 1 Data") #this is to set the working di

genes <- read.csv("QBS103_GSE157103_genes.csv") #this is to rename and retrieve the first gene data fil
the_matrix <- read.csv("QBS103_GSE157103_series_matrix.csv") #this is to rename and retrieve the second

#head(genes) #this is used to visualize the data and only for 6 rows
#head(the_matrix)
```

```
#creating a genes data frame for the genes file
test_genes <- as.data.frame(t(genes))
names(test_genes) <- test_genes[1,] #this allows the genes table to be organized according to names by
test_genes <- test_genes[-1,] #this removes the first row containing "x" in the genes file so that it c
```

```
test_genes$participant_id <- row.names(test_genes) #this will move the participant id into it's own col
combined <- merge(test_genes, the_matrix, by = 'participant_id') #this combines the genes file and matr
```

```
#Histogram using the gene AAAS
library(ggplot2)
setwd("/Users/elodierichard/Documents/QBS103/Project Submission 1 Data")

combined$ABCA13 <- as.numeric(combined$ABCA13) #needed to make them all as.numeric for submission 2
combined$AACS <- as.numeric(combined$AACS)
combined$AAAS <- as.numeric(combined$AAAS) #this is so that the plot can pull just the gene AAAS from t

histogram <- ggplot(combined, aes(x=combined$AAAS)) + #this called on ggplot to use the file "combined"
  geom_histogram(bins = 20, color = 'navy', fill = 'lightblue') +  #this generated the histogram with t
  labs(title = 'Gene Expression of AAAS', #this labeled the title and the axis
       x= 'Gene: AAAS' ,
       y= 'Frequency of AAAS' )
#plot(histogram)
```

```
#Scatterplot of the gene expression of AAAS compared to ferritin levels
library(ggplot2)
combined$ferritin.ng.ml. <- as.numeric(combined$ferritin.ng.ml.) #used to pull out ferritin to plot
```

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

```
#comments mostly the same as for histogram except for a few changes
scatterplot <- ggplot(combined, aes(x= combined$ferritin.ng.ml., y = combined$AAAS)) + #need to specifi
  geom_point(bins = 10, color = 'violet') +  #use geom_point for a scatter plot to be generated
  labs(title = 'Gene Expression of AAAS vs. Ferritin Levels' ,
       x= 'Ferritin Levels (ng/mL)',
       y= 'Gene Expression of AAAS')
```

```
## Warning in geom_point(bins = 10, color = "violet"): Ignoring unknown
## parameters: `bins`
```

```
#for trendline it's geom_smooth
#plot(scatterplot)
```

```
#Scatterplot for Gene Expression vs Age (this was run to compare different data to see differences) not
library(ggplot2)

scatterplot_practice<- ggplot(combined, aes(x= combined$age, y = combined$AAAS)) +
  geom_point(bins = 10, color = 'green') +
  labs(title = 'Gene Expression of AAAS vs. Age' , x= 'Age of Participant (yrs)', y= 'Gene AAAS')
```

```
## Warning in geom_point(bins = 10, color = "green"): Ignoring unknown parameters:
## `bins`
```

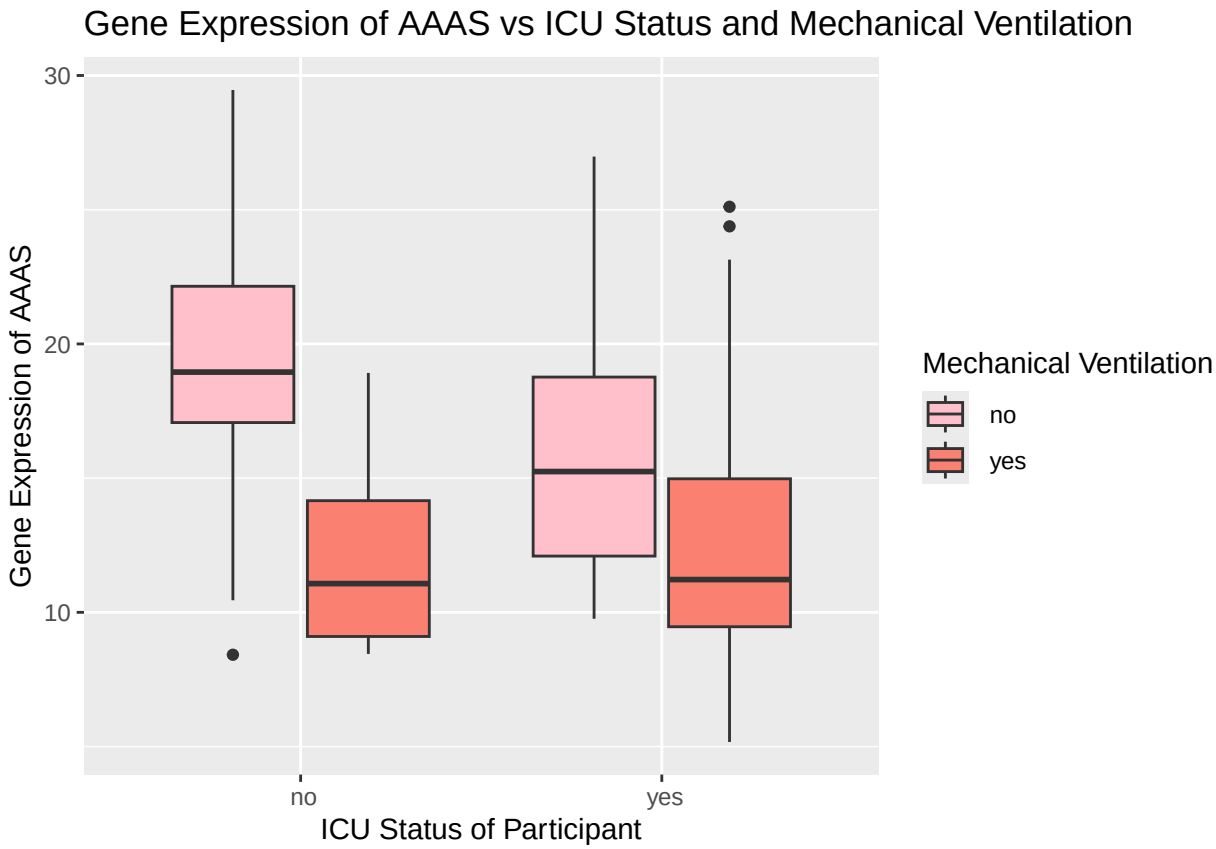
```
#plot(scatterplot_practice)
```

```
#Boxplot comparing gene expression of AAAS related to ICU status depending on Age
library(ggplot2)
#similar process to histogram and scatterplot with a few adjustments
boxplot <- ggplot(combined, aes(x=icu_status, y = AAAS, fill = age)) + #need to add a fill to demonstra
  geom_boxplot(bins = 10, fill = 'maroon') + #to generate a box plot use geom_boxplot
  labs(title = 'Gene Expression of AAAS vs ICU Status and Participant Age',  #to label each attribute o
       x= 'ICU Status of Participant' ,
       y= 'Gene Expression of AAAS',
       fill= 'Age of Participant (yrs)')
```

```
## Warning in geom_boxplot(bins = 10, fill = "maroon"): Ignoring unknown
## parameters: `bins`
```

```
#plot(boxplot)
```

```
### this is the fixed boxplot from the previous submission so that it includes a categorical variable (
#Boxplot comparing gene expression of AAAS related to ICU status depending on Age
library(ggplot2)
#similar process to histogram and scatterplot with a few adjustments
boxplot <- ggplot(combined, aes(x=icu_status, y = AAAS, fill = mechanical_ventilation)) + #need to add
  geom_boxplot() + #to generate a box plot use geom_boxplot
  scale_fill_manual(values = c('pink','salmon')) +
  labs(title = 'Gene Expression of AAAS vs ICU Status and Mechanical Ventilation',  #to label each attr
       x= 'ICU Status of Participant' ,
       y= 'Gene Expression of AAAS',
       fill= 'Mechanical Ventilation')
plot(boxplot)
```



Submission 2 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) Present one of your boxplots in class. Be prepared to explain the gene and covariates you chose and comment on the distribution as if you were presenting your research findings. No slides are required, just bring your plot. In class, be prepared to provide constructive feedback for your classmates (5 pts) Make sure you push your code to your git repository prior to class. As a reminder, we do not need you to share your GitHub repository until the final submission. Pushing this submission to GitHub will be worth 5 pts on the final submission and you can earn 1 additional point on your final project grade if you push 1 extra time along the way (changes between pushes must be significant to earn the extra point).

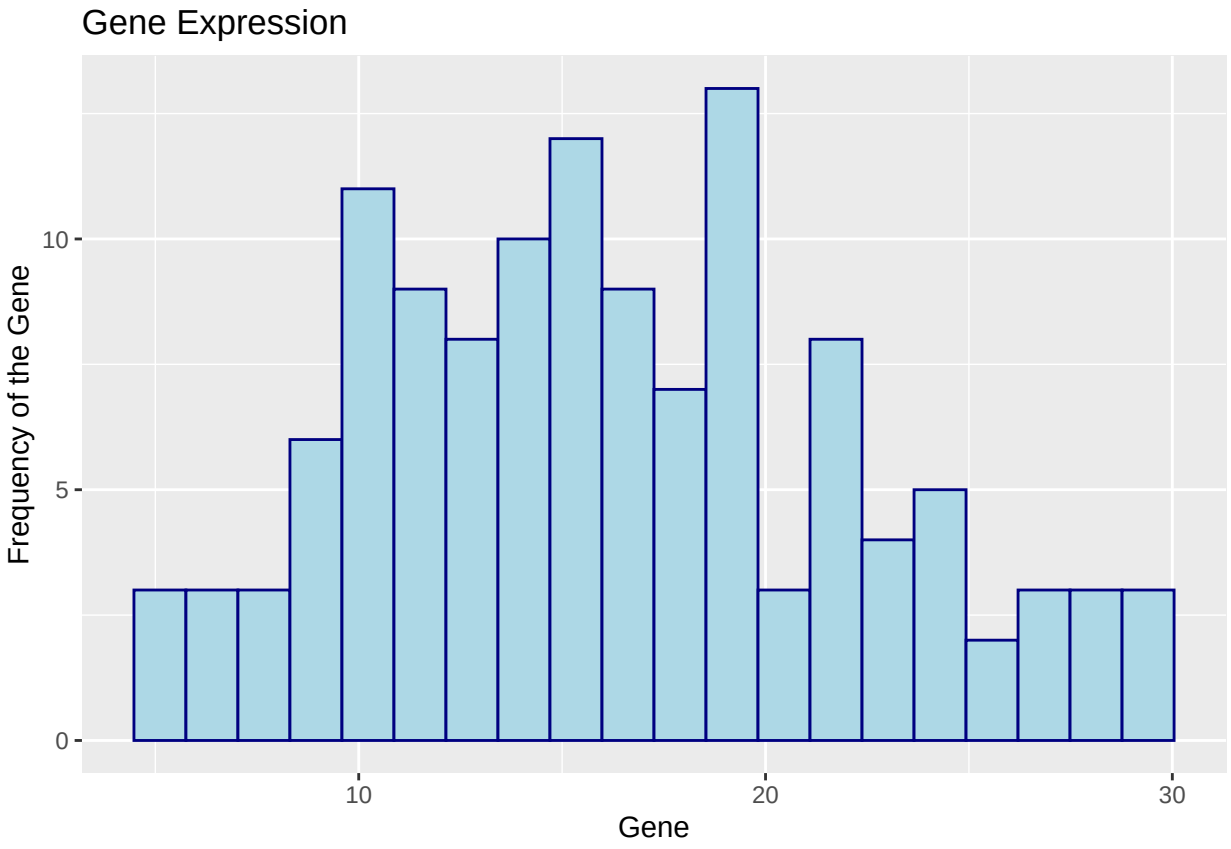
```
## comment out things you don't need anymore

sub_2_plots <- function(data, genes, cat1 , cat2, cont ) { #created a new function with each element de

    histogram2 <- ggplot(data, aes_string(x= genes)) + geom_histogram(bins = 20, color = 'navy', fill =
    scatterplot2 <- ggplot(data, aes_string(x= cont, y = genes)) + geom_point(bins = 10, color = 'viole
   boxplot2 <- ggplot(data, aes_string(x= cat1, y = genes, fill = cat2)) + geom_boxplot() + scale_fill_

    plot(histogram2)
    plot(scatterplot2)
    plot(boxplot2)
}

specific_genes = subset(combined, select = c("AAAS", "AACS", "ABCA13")) #this is to subset the data tab
### had to be sure to make these chosen genes as.numeric so that it would run through
```

```r
for (gene in colnames(specific_genes)) { #created a for loop to run through every gene chosen in the ne
  print(gene)
  print(combined[,gene]) # this will grab each row and the gene columns specifically
  sub_2_plots(combined, gene , cat1 = "icu_status", cat2 = "mechanical_ventilation", cont = "ferritin.ng
```

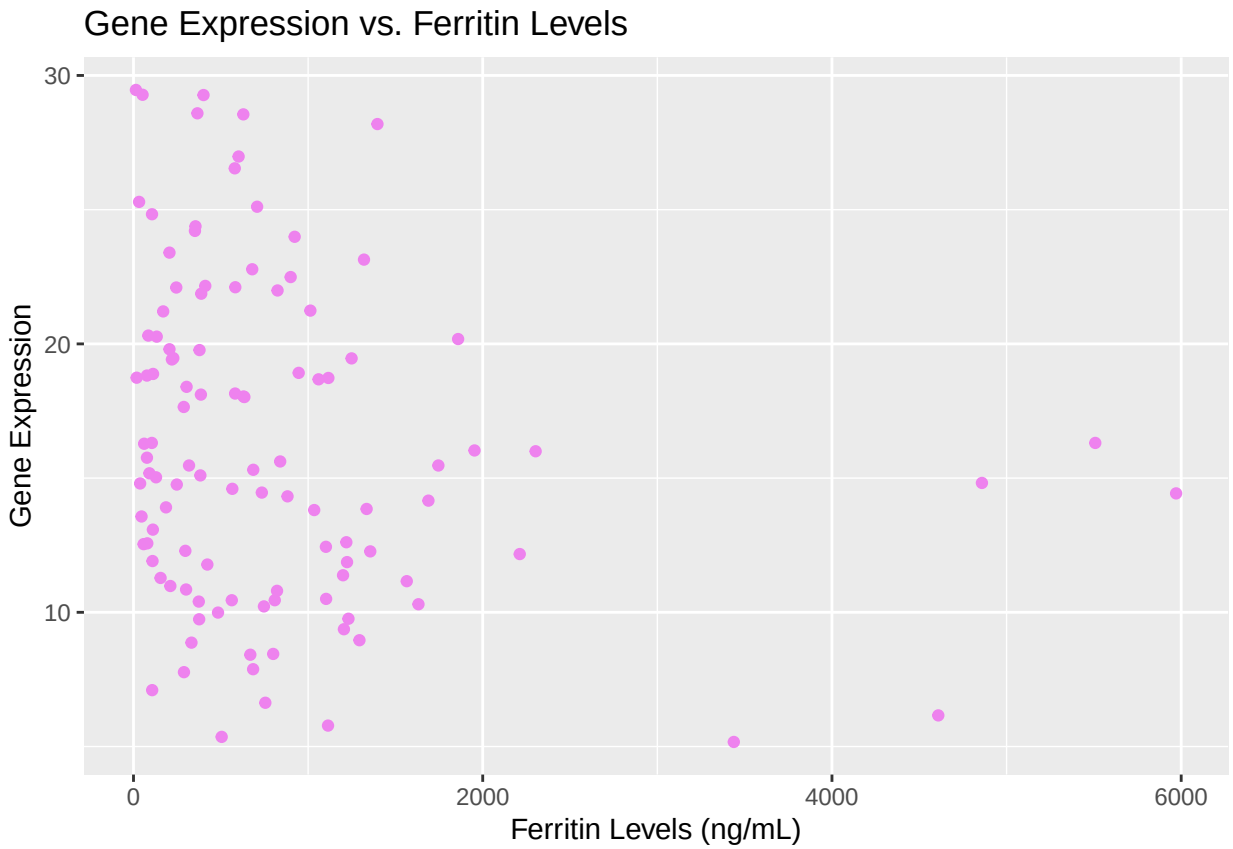
```
## [1] "AAAS"
##   [1] 18.92 18.68 13.85 22.11  8.45 28.59 10.50 22.78 15.47 18.40 26.98  9.10
##  [13]  8.42 29.27 16.00 22.10 10.30  9.37 23.99 19.46 18.82 18.73 12.61  7.10
##  [25]  5.17  8.87 11.16 24.38 15.47 14.32 11.91  9.74 15.31 10.40  8.96 21.24
##  [37] 10.45 14.82 14.16 14.76 12.17 10.22 14.60  6.63 15.10  5.78 10.80  5.36
##  [49] 19.77 12.44 10.85 23.14  6.16 20.18 11.07 16.28 13.81 15.18 25.29 19.47
##  [61] 18.66 21.99 19.80 16.31 15.76  9.99 19.42 28.19 25.11 16.03 23.40 22.49
##  [73] 12.27 29.46 28.55 13.91 14.43  7.88 11.87 18.02 18.88 11.38 17.10 20.27
##  [85] 15.62 11.78 24.21 21.21 14.80 17.65 19.02 13.08 21.87 29.28 18.11 16.89
##  [97] 14.46 18.15  9.76 18.74 12.29 10.45 12.54 15.03 26.54 17.95 13.92 22.16
## [109] 12.57 18.04 11.35 15.53  7.77 24.64 16.26 16.31 10.98 11.28 13.57 24.83
## [121] 17.06 20.31 27.25 21.64  5.54
```

```
## 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 in geom_point(bins = 10, color = "violet"): Ignoring unknown
## parameters: `bins`
```

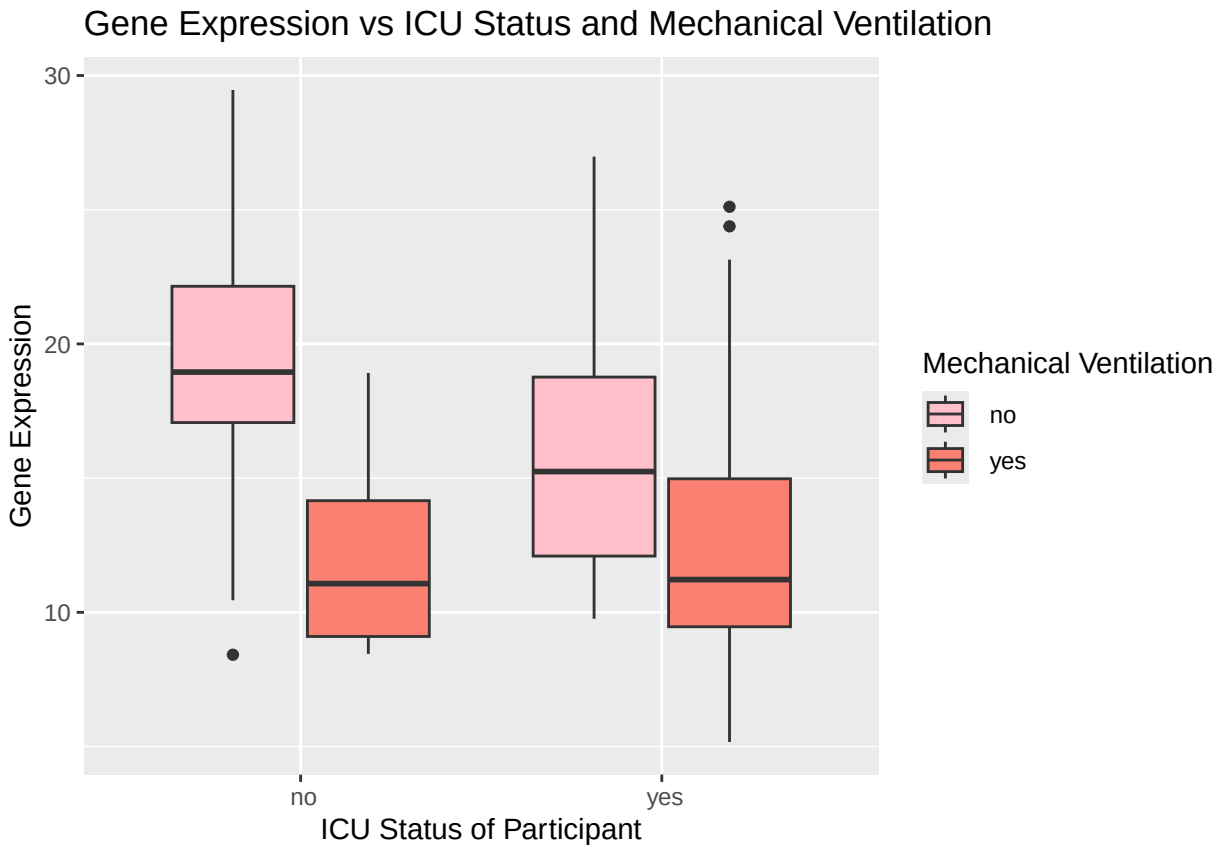


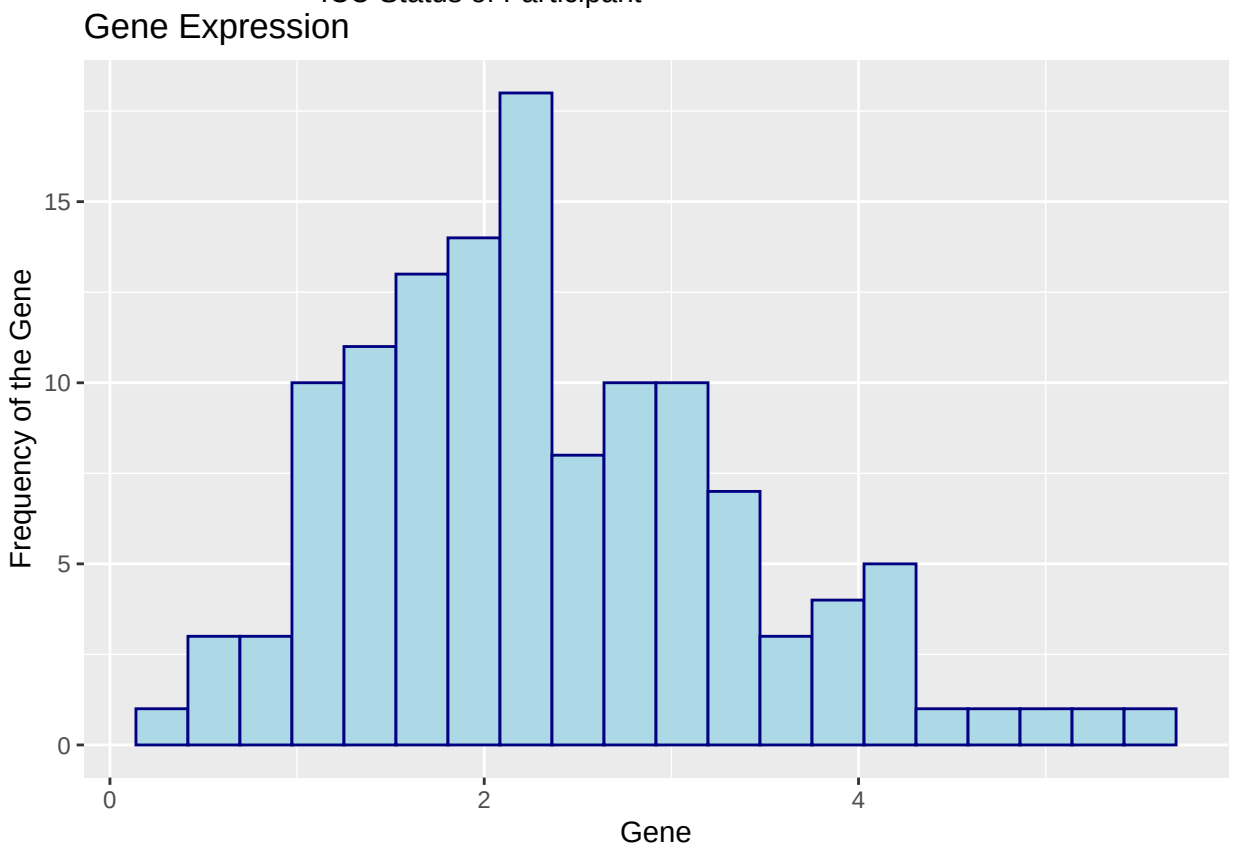
```
## Warning: Removed 16 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
## [1] "AACS"
##   [1] 4.07 3.00 1.83 4.22 1.17 4.24 2.10 4.86 2.90 1.84 2.79 1.06 0.64 3.90 3.61
##  [16] 2.73 2.16 2.94 3.40 4.32 2.89 2.33 1.94 1.11 1.05 1.45 2.02 3.92 2.22 2.97
##  [31] 1.65 1.57 1.56 1.74 1.88 2.89 1.95 2.66 2.52 1.80 1.12 0.55 2.47 1.21 2.23
##  [46] 1.44 2.03 1.26 1.81 2.03 1.48 2.38 0.62 2.22 2.18 2.52 1.33 2.09 5.63 2.20
##  [61] 3.25 4.23 3.67 2.27 1.78 2.14 2.98 3.04 3.19 2.09 4.18 3.00 1.87 5.35 5.08
##  [76] 2.25 1.78 0.83 1.23 2.42 2.18 1.47 2.31 3.02 1.74 1.77 2.74 3.83 1.67 2.52
##  [91] 3.25 2.26 3.83 3.29 3.25 2.78 1.49 2.78 1.11 2.63 1.07 0.95 2.24 1.27 3.05
## [106] 1.73 1.91 3.22 2.31 3.11 1.02 1.48 0.90 3.57 1.81 2.47 1.60 1.30 2.00 2.74
## [121] 1.96 1.54 3.42 1.43 0.35
```

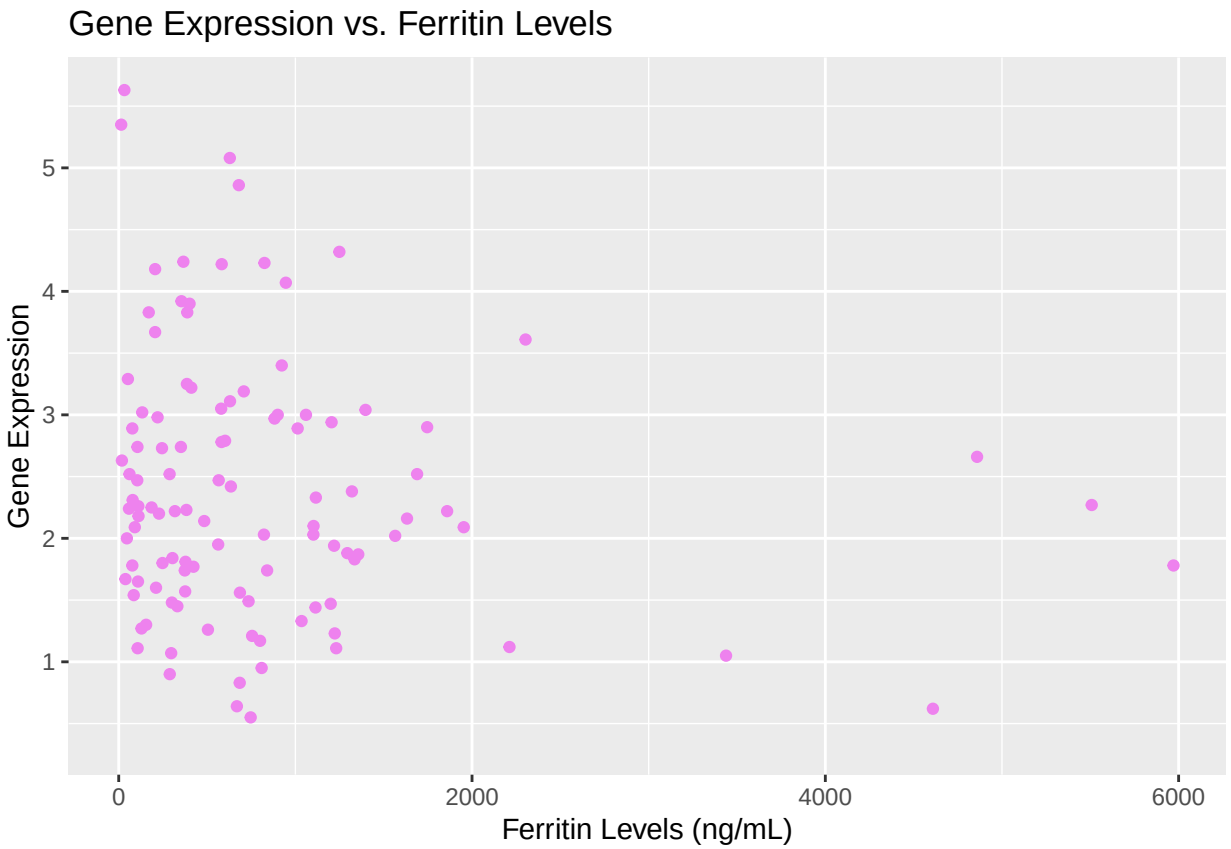
```
## Warning in geom_point(bins = 10, color = "violet"): Ignoring unknown
## parameters: `bins`
```





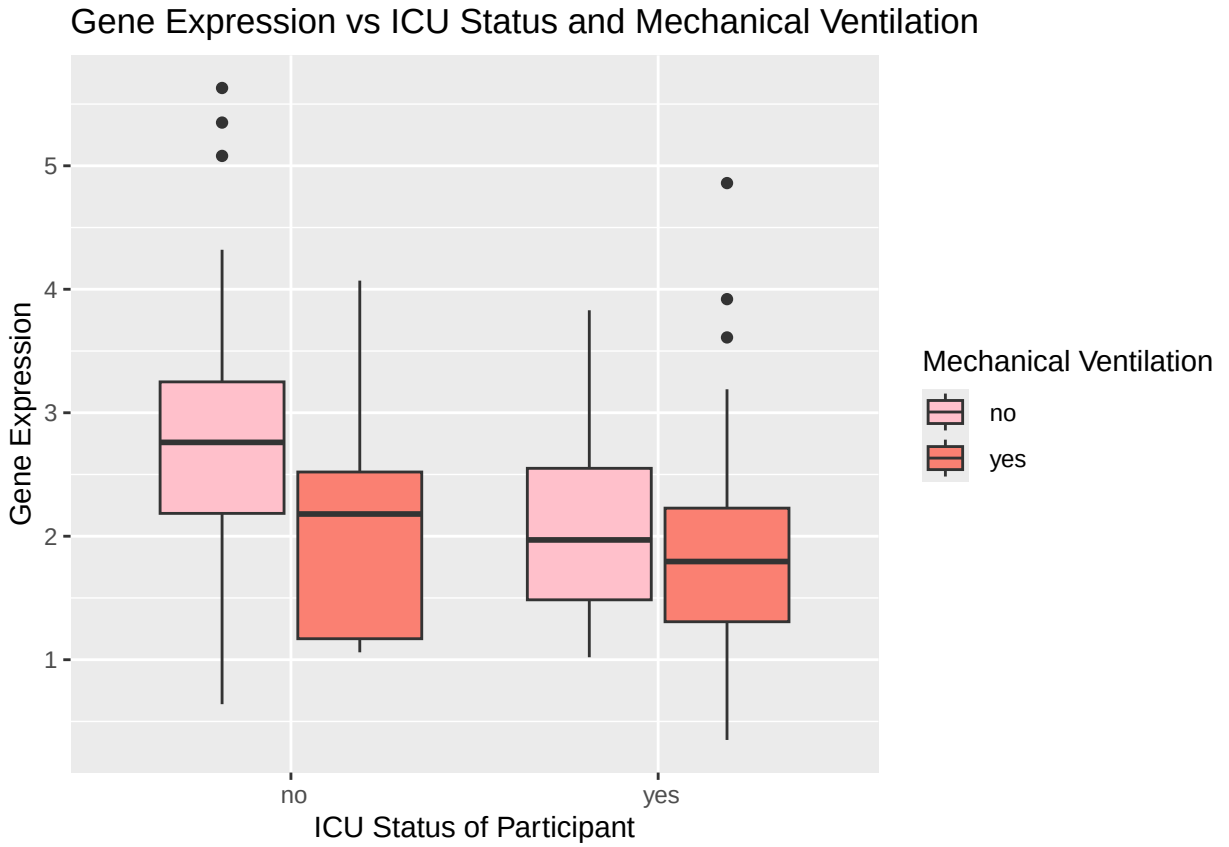

```
## Warning: Removed 16 rows containing missing values or values outside the scale range
```

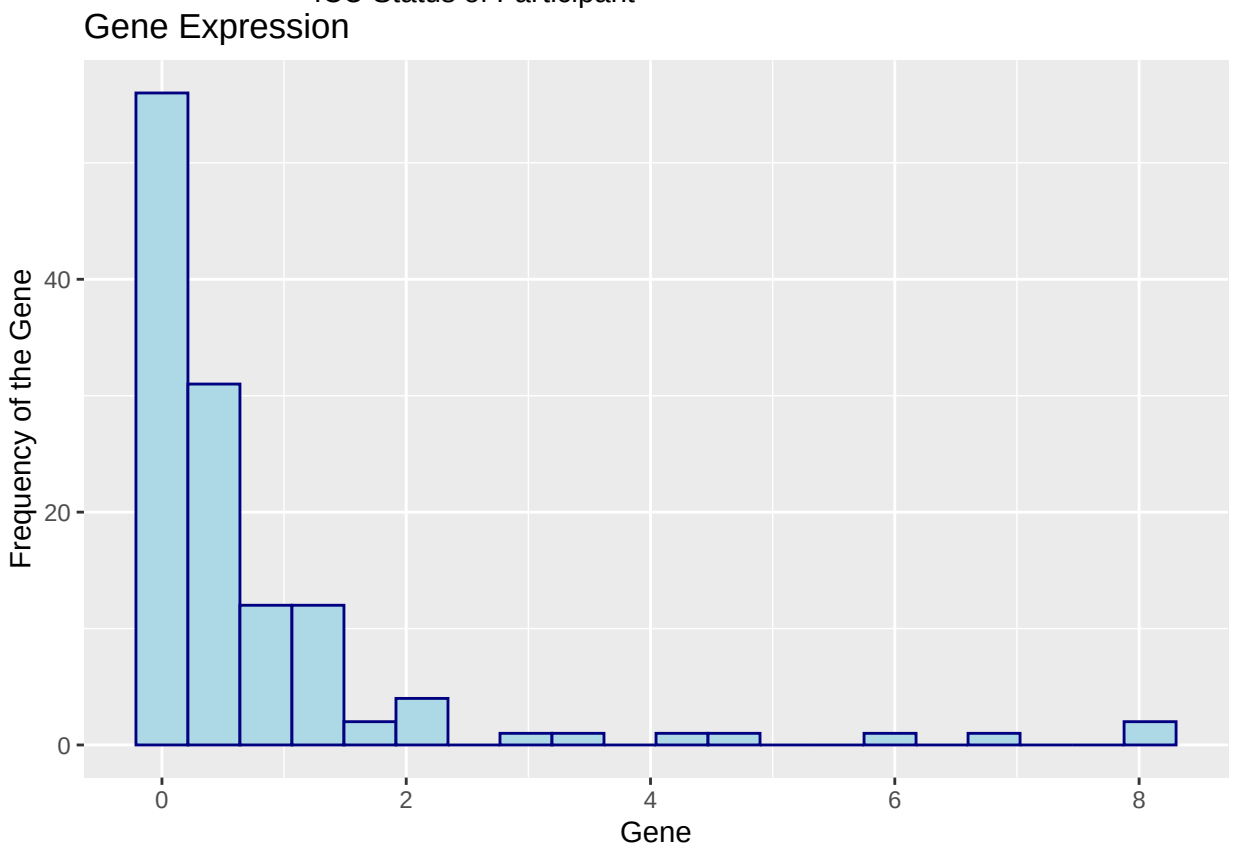
```
## (`geom_point()`).
```




```
## [1] "ABCA13"
##   [1] 0.49 3.36 0.26 0.13 0.16 0.23 6.62 1.29 8.10 0.31 0.64 0.40 0.09 0.07 4.78
##  [16] 0.11 0.62 1.05 1.24 1.09 1.02 0.14 1.42 0.40 3.01 1.36 0.94 5.79 0.22 2.15
##  [31] 0.37 0.19 0.42 1.59 4.08 0.51 0.45 1.21 0.99 0.17 0.07 1.01 1.37 0.19 1.26
##  [46] 0.95 0.48 0.07 0.01 0.24 1.19 0.24 0.56 1.31 0.07 2.08 2.34 0.83 0.51 0.08
##  [61] 0.08 0.37 0.16 1.43 0.26 0.80 0.25 0.19 0.10 0.72 0.32 0.13 7.90 0.05 0.15
##  [76] 2.15 0.21 0.28 0.70 0.09 0.13 0.53 0.04 0.58 0.37 0.01 0.08 0.49 0.50 0.53
##  [91] 0.04 1.77 0.04 0.01 0.22 0.02 0.09 1.22 0.05 0.03 0.26 0.14 0.07 0.01 0.01
## [106] 0.01 0.18 0.11 0.34 0.10 0.02 0.09 0.03 0.02 0.77 0.02 0.12 0.05 0.13 0.04
## [121] 0.01 0.01 0.07 0.03 0.06
```

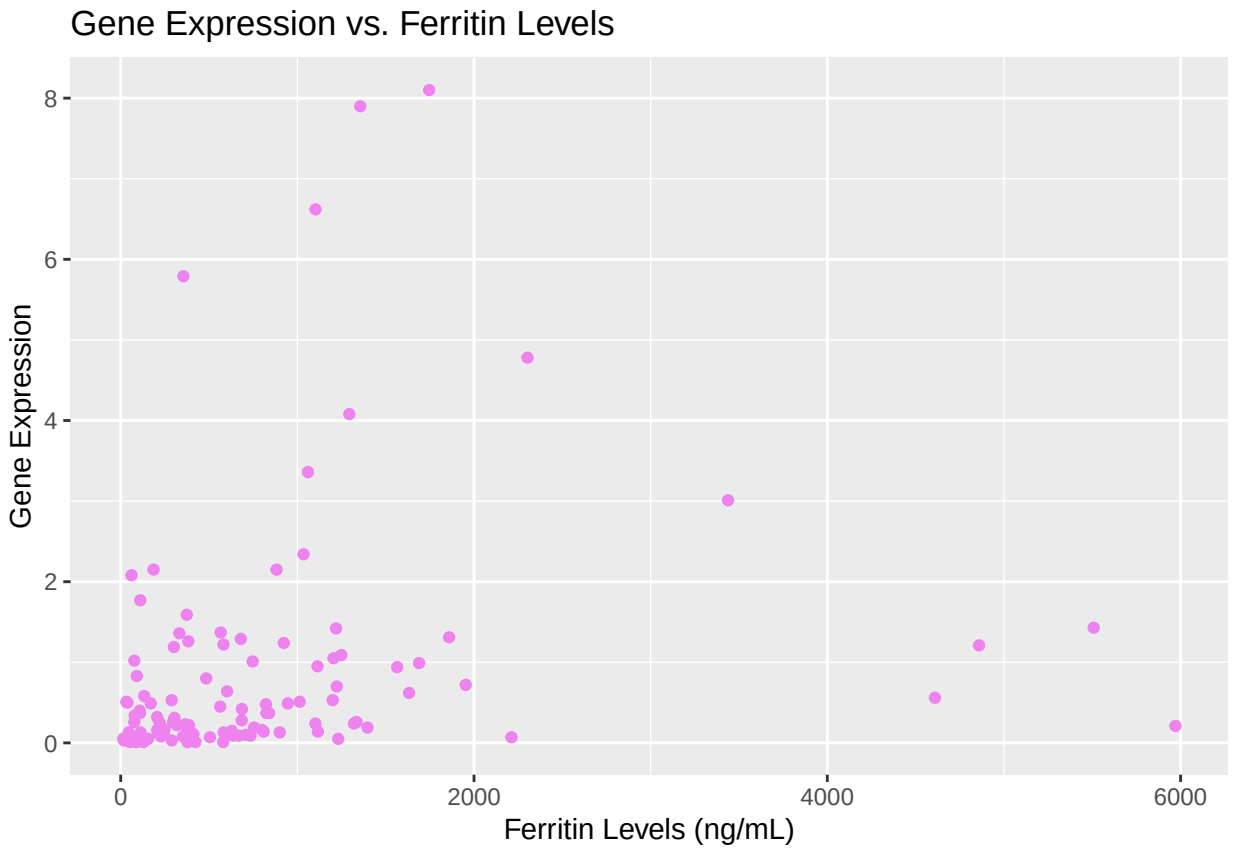
```
## Warning in geom_point(bins = 10, color = "violet"): Ignoring unknown
## parameters: `bins`
```

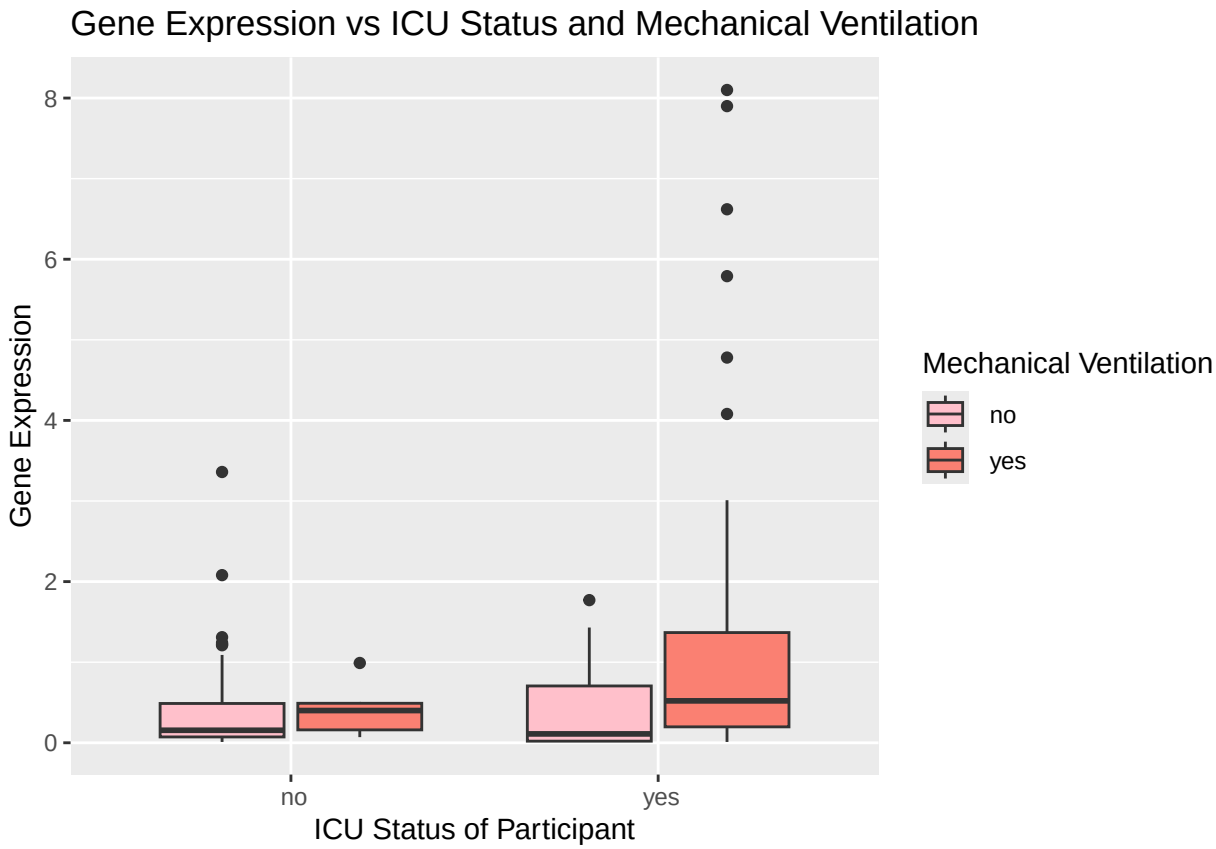





```
## Warning: Removed 16 rows containing missing values or values outside the scale range
```

```
## (`geom_point()`).
```


Gene Expression vs. Ferritin Levels
Gene Expression
0
2
4
6
8
0
2000
4000
6000
Ferritin Levels (ng/mL)




```
# sub_2_plots(combined, "AAAS", cat1 = "icu_status", cat2 = "mechanical_ventilation", cont = "ferritin
```

‘‘‘