QBS103 Submission 2

Isabelle Kressy

2025-07-19

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
# this is copied from my submission 1


# set working directory
setwd("~/Desktop/Dartmouth/Foundations of Data Science")

# import csv files
genes_df <- read_csv(file = 'QBS103_GSE157103_genes.csv')
```

```
## New names:
## Rows: 100 Columns: 127
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): ...1 dbl (126): COVID_01_39y_male_NonICU, COVID_02_63y_male_NonICU,
## COVID_03_33y_...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

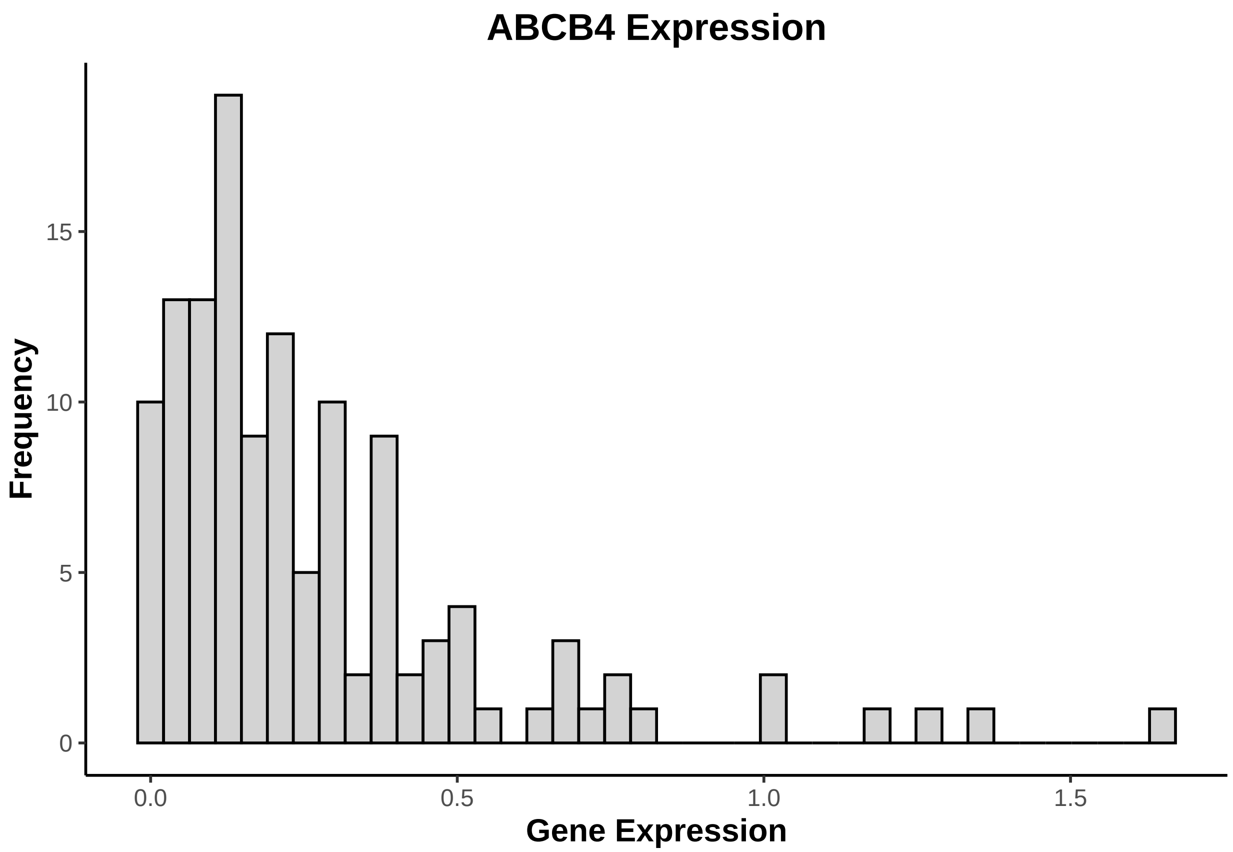
```r
meta_df <- read_csv(file = 'QBS103_GSE157103_series_matrix-1.csv')
```

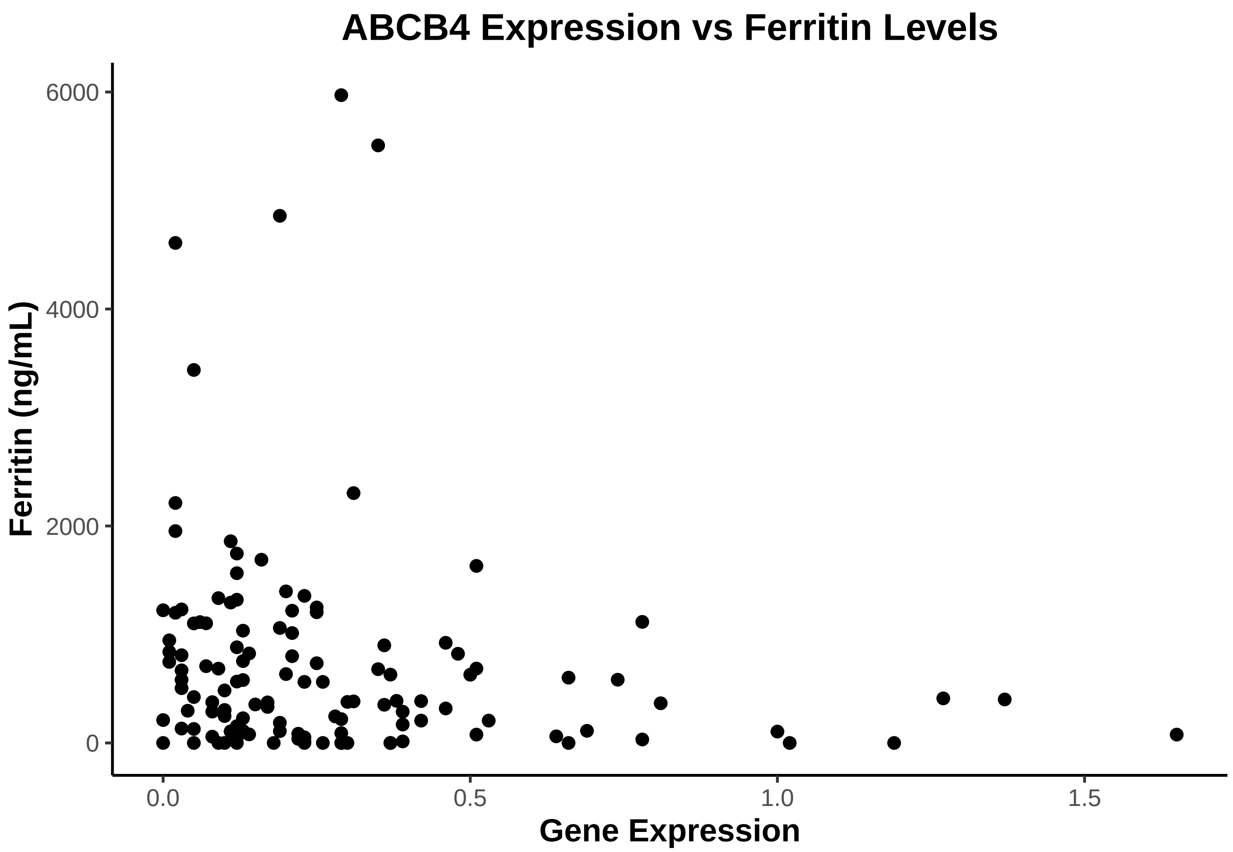
```
## Rows: 126 Columns: 25
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (21): participant_id, geo_accession, status, !Sample_submission_date, la...
## dbl  (4): channel_count, charlson_score, ventilator-free_days, hospital-free...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

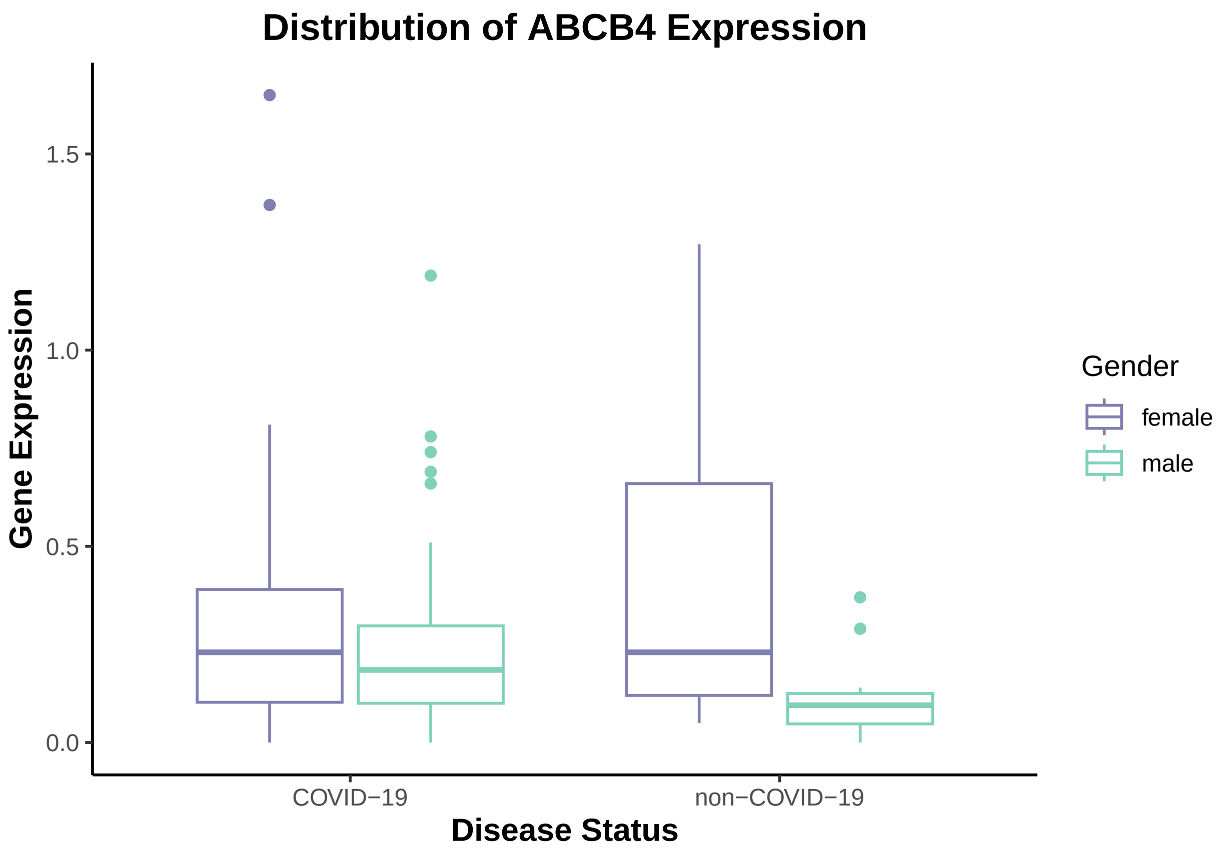
```r
# combine dataframes and subset for plotting

## tidy genes_df to a format for df merging (ie change format, change column names)
genes_df_tidy <- genes_df %>%
  pivot_longer(cols = c(-'...1'), names_to = 'participant_id2',
               values_to = 'Gene Expression') %>%
  mutate('Gene' = ...1) %>%
  select(-'...1') %>%
  pivot_wider(values_from = 'Gene Expression', names_from = 'Gene')

## join data frames by participant_id
full_df <- cbind(genes_df_tidy, meta_df)
full_df$`ferritin(ng/ml)`[full_df$`ferritin(ng/ml)` == "unknown"] <- 0
full_df$`ferritin(ng/ml)` <- as.numeric(full_df$`ferritin(ng/ml)`)

## create factors of sex and disease_status for correct plotting
full_df$sex <- factor(full_df$sex, levels = c('female', 'male'))
full_df$disease_status <- factor(full_df$disease_status,
                                 levels = c('disease state: COVID-19',
                                            'disease state: non-COVID-19'))

# gene = ABCB4, AAMP, AASS
# categorical covariates = sex and disease status
# continuous covariate = ferritin(ng/ml)
```

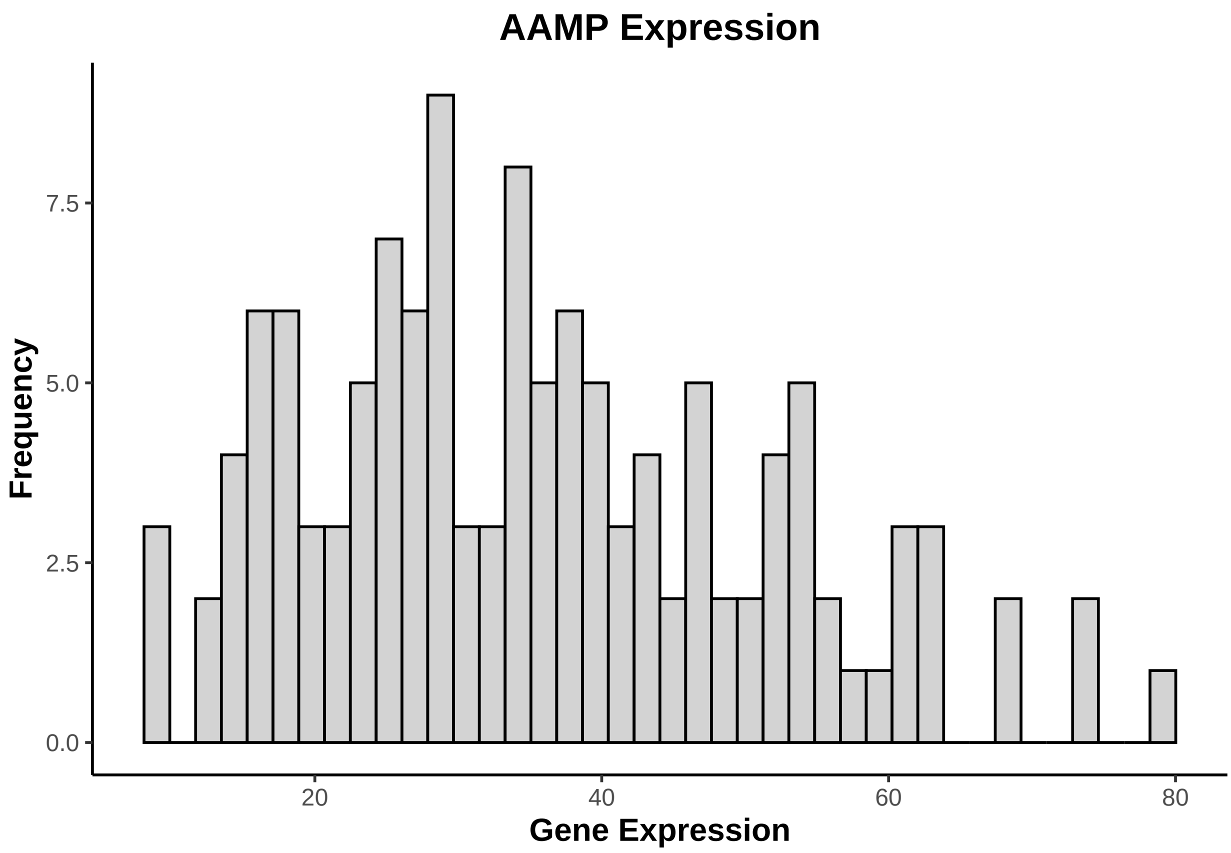
```r
# define inputs for function
genes <- list('ABCB4', 'AAMP', 'AASS')
continuous_covariate <- list('ferritin(ng/ml)')
categorical_covariate <- list('sex', 'disease_status')
columns <- colnames(full_df)

# histogram, scatterplot, boxplot theme
myTheme <- theme(panel.border = element_blank(),
                 panel.grid.major = element_blank(),
                 panel.grid.minor = element_blank(),
                 axis.line = element_line(colour = "black"),
                 plot.background = element_blank(),
                 panel.background = element_blank(),
                 plot.title = element_text(size = 14, face = 'bold', hjust = 0.5),
                 axis.title.x = element_text(size = 12, face = 'bold'),
                 axis.title.y = element_text(size = 12, face = 'bold'))
```

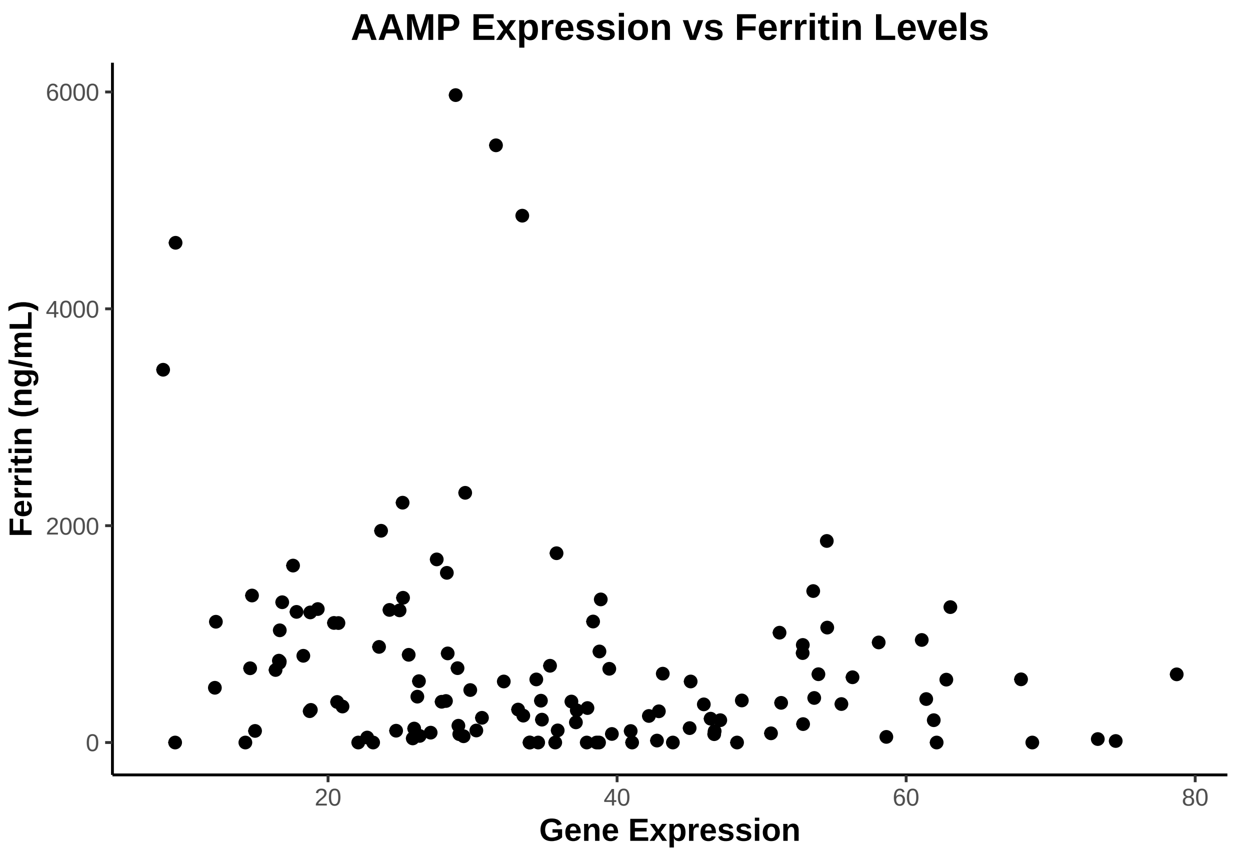
```r
# create function for plotting
# found the paste() function in chat to insert an object into a string
# used chat to get rid of NA values in sex for boxplot plotting (see data = ... part)
plotting_function <- function(full_df,genes,continuous_covariate,categorical_covariate){
  for (gene in genes){
    hist_plot <- ggplot(data = full_df, aes(x = .data[[gene]])) +
      geom_histogram(bins = 40, fill = 'lightgrey', color = 'black') +
      labs(title = paste(gene, 'Expression'), x = 'Gene Expression', y = 'Frequency') +
      myTheme
  print(hist_plot)

    for (cont_var in continuous_covariate){
      scat_plot <- ggplot(data = full_df, aes(x = .data[[gene]], y = .data[[cont_var]])) +
        geom_point(size = 1.75) +
        labs(title = paste(gene, 'Expression vs Ferritin Levels'),
            x = 'Gene Expression', y = 'Ferritin (ng/mL)') +
        myTheme
      print(scat_plot)

      bp <- ggplot(data = subset(full_df, !is.na(sex)),
                   aes(x = .data[[categorical_covariate[[2]]]],
                       y = .data[[gene]], color = .data[[categorical_covariate[[1]]]])) +
        geom_boxplot() +
        scale_color_manual(values = c('#7F80B1', '#7FD1B9', 'black')) +
        labs(x = 'Disease Status', y = 'Gene Expression',
             title = paste('Distribution of', gene, 'Expression'),
             color = 'Gender') +
        scale_x_discrete(labels = c('COVID-19', 'non-COVID-19')) +
        myTheme
  print(bp)

  }
  }

}

plotting_function(full_df = full_df, genes = genes, continuous_covariate = continuous_covariate,
                  categorical_covariate = categorical_covariate)
```

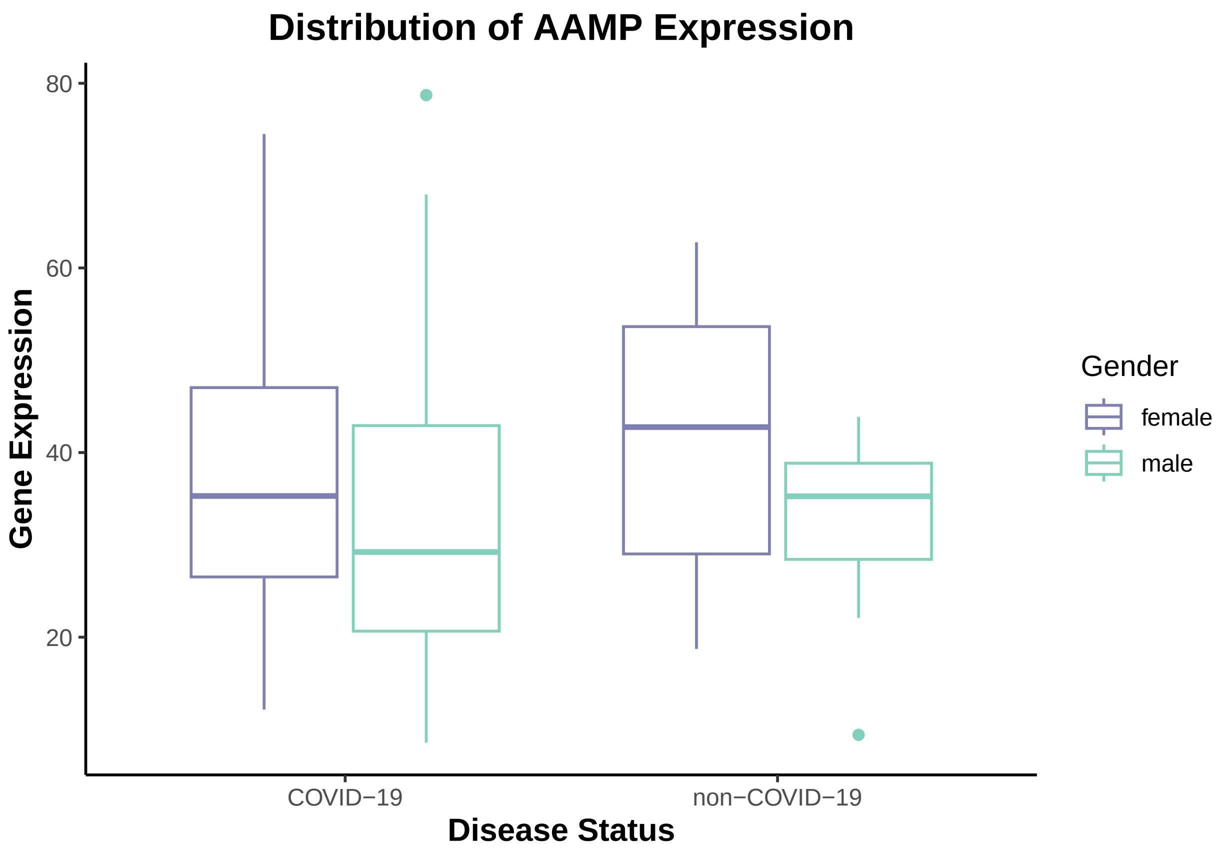

ABCB4 Expression
Frequency
15
10
5
0
0.0
0.5
1.0
1.5
Gene Expression

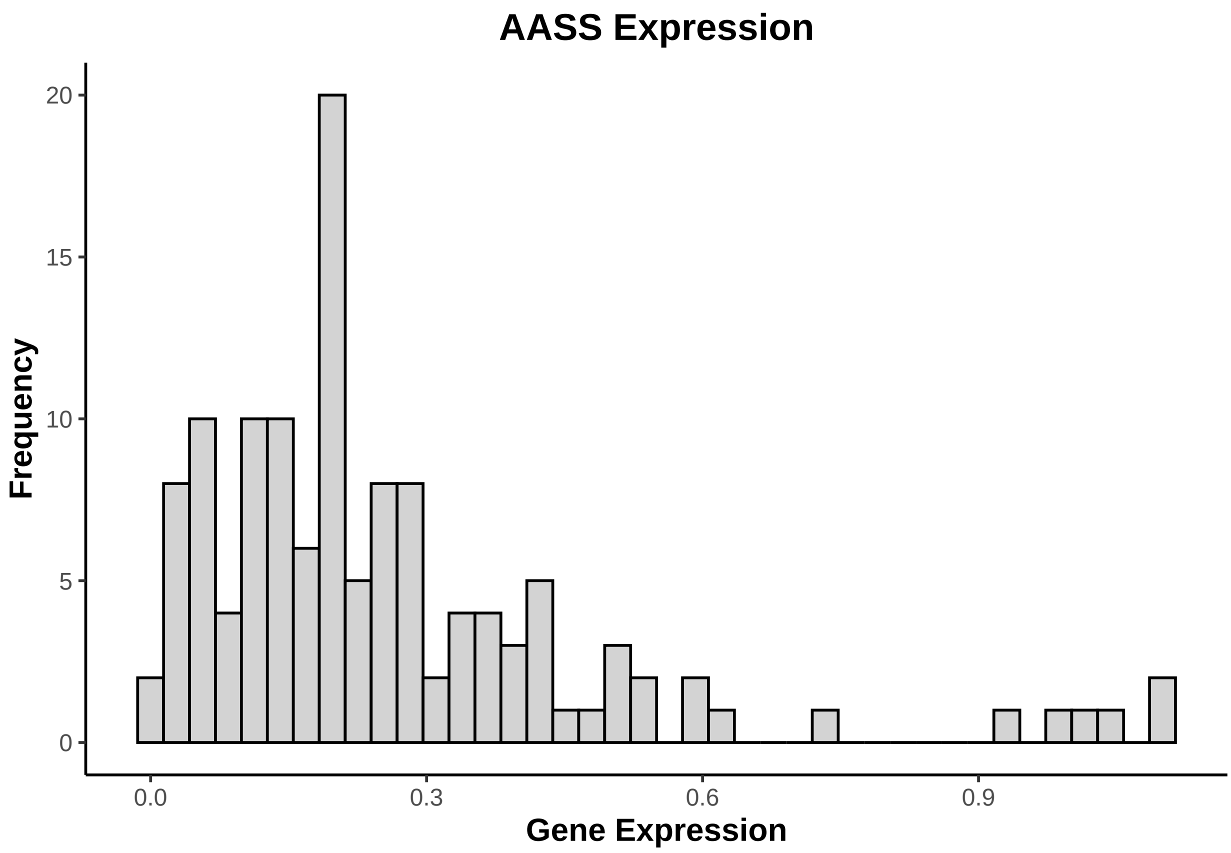

ABCB4 Expression vs Ferritin Levels
Ferritin (ng/mL)
6000
4000
2000
0
0.0
0.5
1.0
1.5
Gene Expression

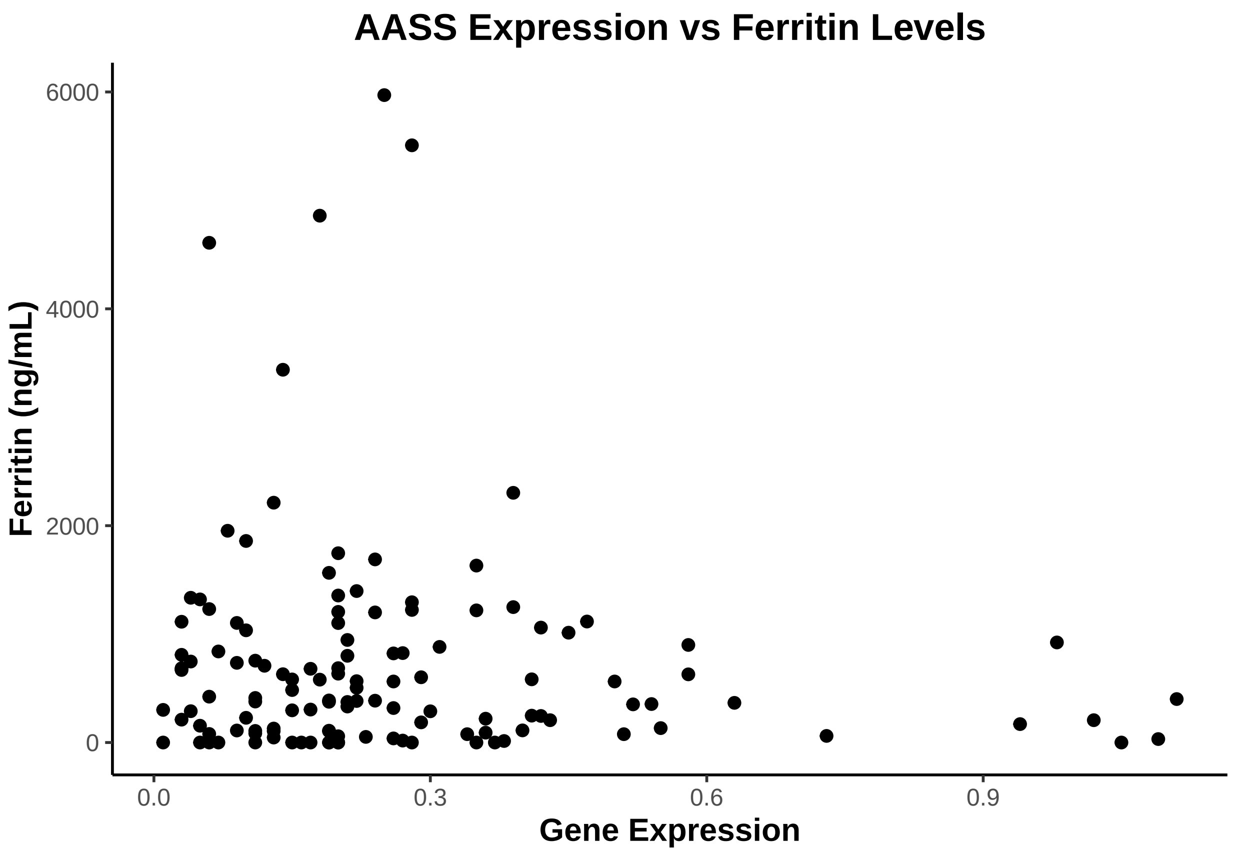

Distribution of ABCB4 Expression
Gene Expression
1.5
1.0
0.5
0.0
COVID−19
non−COVID−19
Disease Status
Gender
female
male

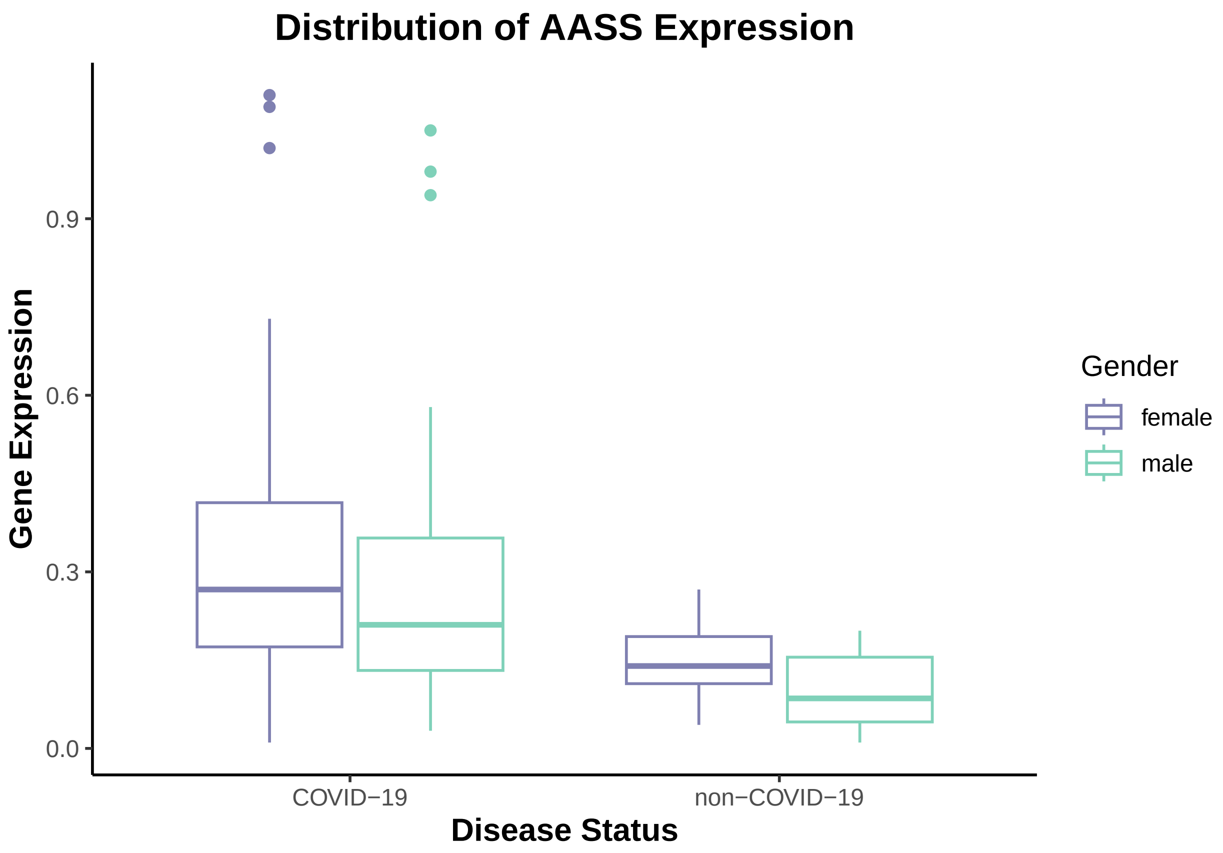

AAMP Expression
Frequency
7.5
5.0
2.5
0.0
20
40
60
80
Gene Expression


AAMP Expression vs Ferritin Levels
Ferritin (ng/mL)
6000
4000
2000
0
20
40
60
80
Gene Expression


Distribution of AAMP Expression
Gene Expression
80
60
40
20
COVID−19
non−COVID−19
Disease Status
Gender
female
male


AASS Expression
Frequency
20
15
10
5
0
0.0
0.3
0.6
0.9
Gene Expression


AASS Expression vs Ferritin Levels
Ferritin (ng/mL)
6000
4000
2000
0
0.0
0.3
0.6
0.9
Gene Expression


Distribution of AASS Expression
Gene Expression
0.0
0.3
0.6
0.9
COVID−19
non−COVID−19
Disease Status
Gender
female
male