QBS103 Submission 1

2025-07-12

and will need to be

and will need to be linked.

```
# call packages
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 errors
```

```
# 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`
```

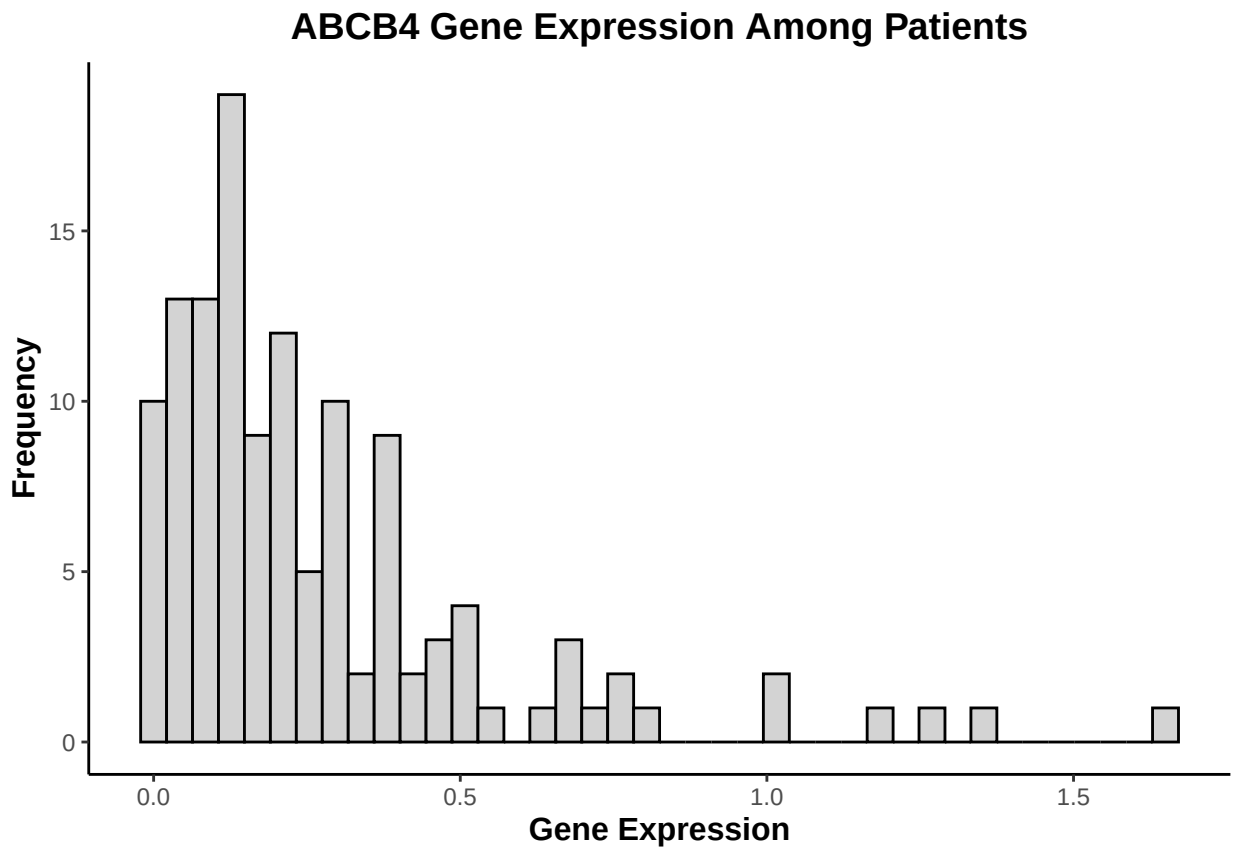
```
meta_df <- read_csv(file = 'QBS103_GSE157103_series_matrix-1.csv')
```

```
## 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.
```

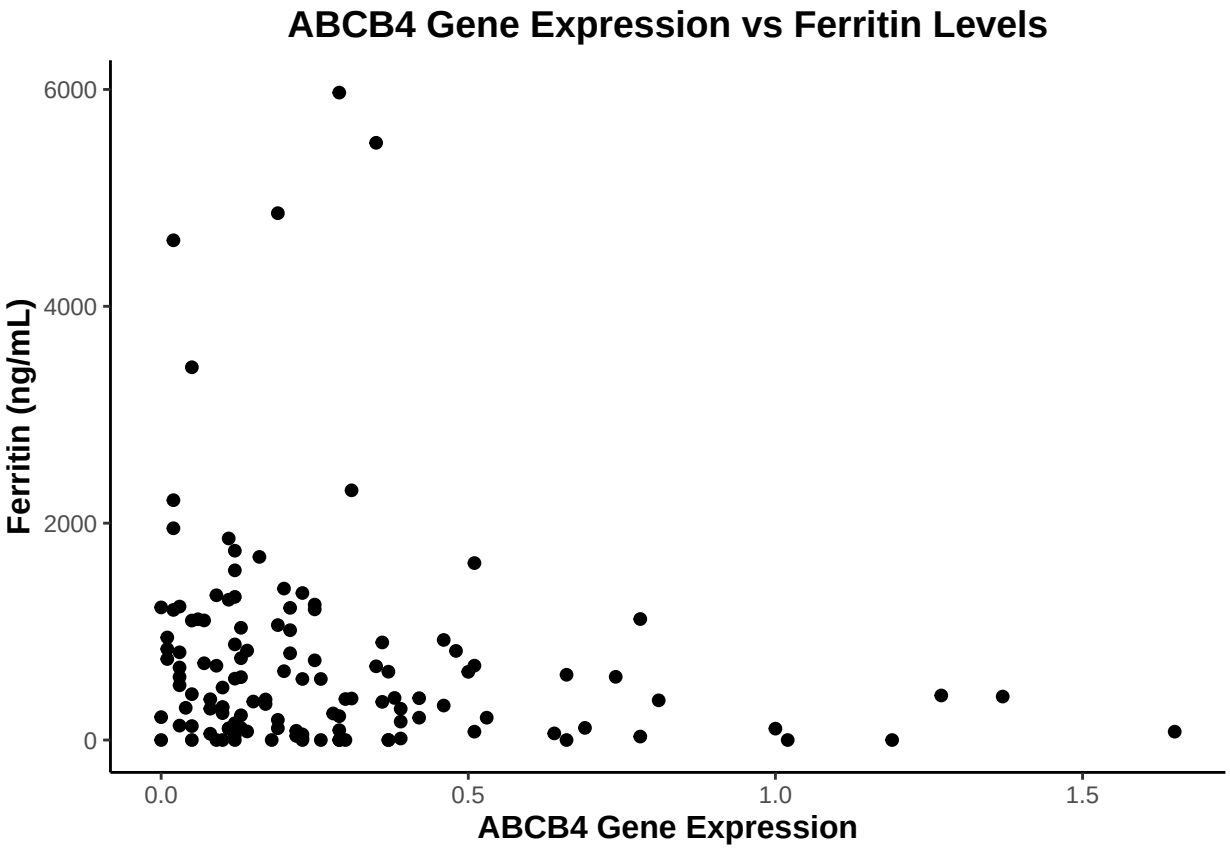
```r
# combine dataframes and subset for plotting

## 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_id', 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)
```

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

# subset full_df for gene of interest and covariates
sub1_df <- full_df %>%
  select(c('participant_id', 'ABCB4', 'disease_status', 'sex', 'ferritin(ng/ml)'))


# replace unknown values with 0, used chat for this
# the brackets are indexing the column ferritin(ng/ml) in sub1_df
# searching for the values equal to unknown, and replacing them with 0 (setting their new value to 0)
sub1_df$`ferritin(ng/ml)`[sub1_df$`ferritin(ng/ml)` == "unknown"] <- 0


# convert ferritin column values to a numeric class for plotting
# class(sub1_df$`ferritin(ng/ml)`) - used this to check the class before plotting
sub1_df$`ferritin(ng/ml)` <- as.numeric(sub1_df$`ferritin(ng/ml)`)


# define theme for plots
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'))
```

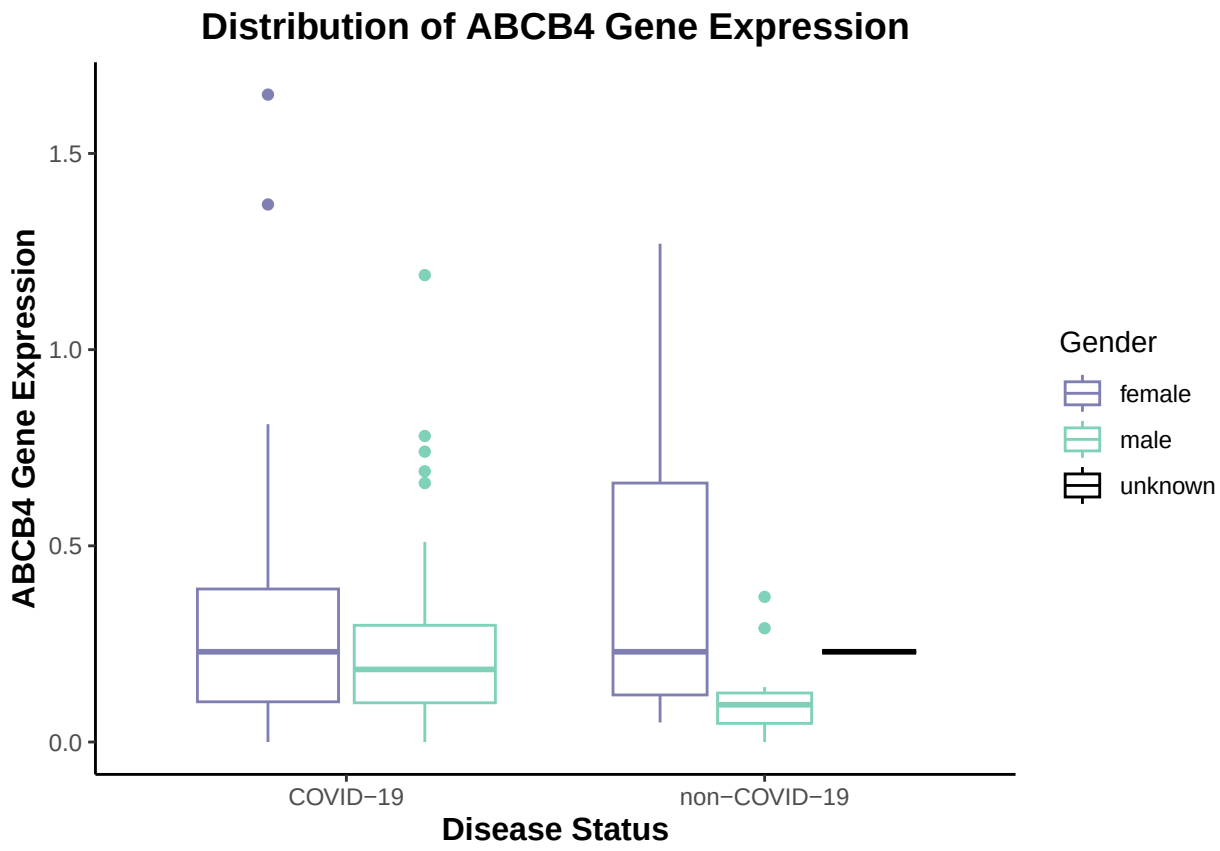
```r
# histogram for gene expression, used the help() function for aesthetics
ggplot(data = sub1_df, aes(x = ABCB4)) +
  geom_histogram(bins = 40, fill = 'lightgrey', color = 'black') +
  labs(title = 'ABCB4 Gene Expression Among Patients', x = 'Gene Expression', y = 'Frequency') +
  myTheme
```




```
# scatterplot for gene expression and continuous covariate
ggplot(data = sub1_df, aes(x = ABCB4, y = `ferritin(ng/ml)`)) +
  geom_point(size = 1.75) +
  labs(title = 'ABCB4 Gene Expression vs Ferritin Levels',
       x = 'ABCB4 Gene Expression', y = 'Ferritin (ng/mL)') +
  myTheme
```




```
# boxplot of gene expression separated by both categorical covariates
# included unknown just in case because i'm not sure if it was genuinely unknown or non-binary folks
ggplot(sub1_df,aes(x = disease_status, y = ABCB4, color = sex)) +
  geom_boxplot() +
  scale_color_manual(values = c('#7F80B1', '#7FD1B9', 'black')) +
  labs(x = 'Disease Status', y = 'ABCB4 Gene Expression', title = 'Distribution of ABCB4 Gene Expressio
  scale_x_discrete(labels = c('COVID-19', 'non-COVID-19')) +
  myTheme
```


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