QBS 103 Final Project V2

Elizabeth Chin

2023-08-05

R Markdown

```
#set working directory
setwd("/Users/li_li/desktop/QBS_103_Final_Project/Data")

#read in csv files
gene.exp <- read.csv("QBS103_finalProject_geneExpression.csv", row.names = 1)

meta.data <- read.csv("QBS103_finalProject_metadata.csv", row.names = 1)
```

```
#subset gene expression ABCA1, ABCA2, ABCA3
gene.ABCA123 <- gene.exp[c('ABCA1','ABCA2','ABCA3'),]

#transform rows to columns
gene.ABCA123 <- as.data.frame(t(gene.ABCA123))
head(gene.ABCA123)
```

```
##                          ABCA1 ABCA2 ABCA3
## COVID_01_39y_male_NonICU 32.30  8.47  0.37
## COVID_02_63y_male_NonICU 15.84  9.49  0.71
## COVID_03_33y_male_NonICU 34.38 14.24  0.17
## COVID_04_49y_male_NonICU 14.24  6.37  0.94
## COVID_05_49y_male_NonICU 18.39  5.90  0.17
## COVID_06_.y_male_NonICU   3.64  6.18  0.43
```

```
#column bind two data frames
ABCA123.gene.exp.meta <- cbind(gene.ABCA123,meta.data)
head(ABCA123.gene.exp.meta, 1)
```

```
##                          ABCA1 ABCA2 ABCA3 geo_accession                status
## COVID_01_39y_male_NonICU  32.3  8.47  0.37    GSM4753021 Public on Aug 29 2020
##                          X.Sample_submission_date last_update_date type
## COVID_01_39y_male_NonICU              Aug 28 2020      Aug 29 2020  SRA
##                          channel_count             source_name_ch1 organism_ch1
## COVID_01_39y_male_NonICU             1 Leukocytes from whole blood Homo sapiens
##                                    disease_status age   sex icu_status apacheii
## COVID_01_39y_male_NonICU disease state: COVID-19  39  male         no       15
##                          charlson_score mechanical_ventilation
## COVID_01_39y_male_NonICU              0                    yes
```

```
##                           ventilator.free_days
## COVID_01_39y_male_NonICU                    0
##                           hospital.free_days_post_45_day_followup
## COVID_01_39y_male_NonICU                                        0
##                           ferritin.ng.ml. crp.mg.l. ddimer.mg.l_feu.
## COVID_01_39y_male_NonICU              946      73.1               1.3
##                           procalcitonin.ng.ml.. lactate.mmol.l. fibrinogen sofa
## COVID_01_39y_male_NonICU                     36             0.9        513    8
```

```
#convert age to an integer
ABCA123.gene.exp.meta$age <- as.integer(ABCA123.gene.exp.meta$age)
```

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

```
head(ABCA123.gene.exp.meta, 1)
```

```
##                           ABCA1 ABCA2 ABCA3 geo_accession                status
## COVID_01_39y_male_NonICU  32.3  8.47  0.37    GSM4753021 Public on Aug 29 2020
##                           X.Sample_submission_date last_update_date type
## COVID_01_39y_male_NonICU              Aug 28 2020      Aug 29 2020  SRA
##                           channel_count             source_name_ch1 organism_ch1
## COVID_01_39y_male_NonICU              1 Leukocytes from whole blood Homo sapiens
##                                     disease_status age   sex icu_status apacheii
## COVID_01_39y_male_NonICU disease state: COVID-19  39  male         no       15
##                           charlson_score mechanical_ventilation
## COVID_01_39y_male_NonICU              0                     yes
##                           ventilator.free_days
## COVID_01_39y_male_NonICU                    0
##                           hospital.free_days_post_45_day_followup
## COVID_01_39y_male_NonICU                                        0
##                           ferritin.ng.ml. crp.mg.l. ddimer.mg.l_feu.
## COVID_01_39y_male_NonICU              946      73.1               1.3
##                           procalcitonin.ng.ml.. lactate.mmol.l. fibrinogen sofa
## COVID_01_39y_male_NonICU                     36             0.9        513    8
```

```
#create a new data frame to filter the unknown in sex
library(tidyverse)
```

```
## -- 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 errors
```

```
clean_data <- ABCA123.gene.exp.meta %>%
  filter(!grepl('unknown', sex))
```

```r
#create theme
newBlankTheme <- theme(# Remove all the extra 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
        plot.background = element_rect(fill = "white"),
        panel.background = element_blank(),
        legend.key = element_rect(fill = 'white'),
        # Move legend
        legend.position = 'top')
```

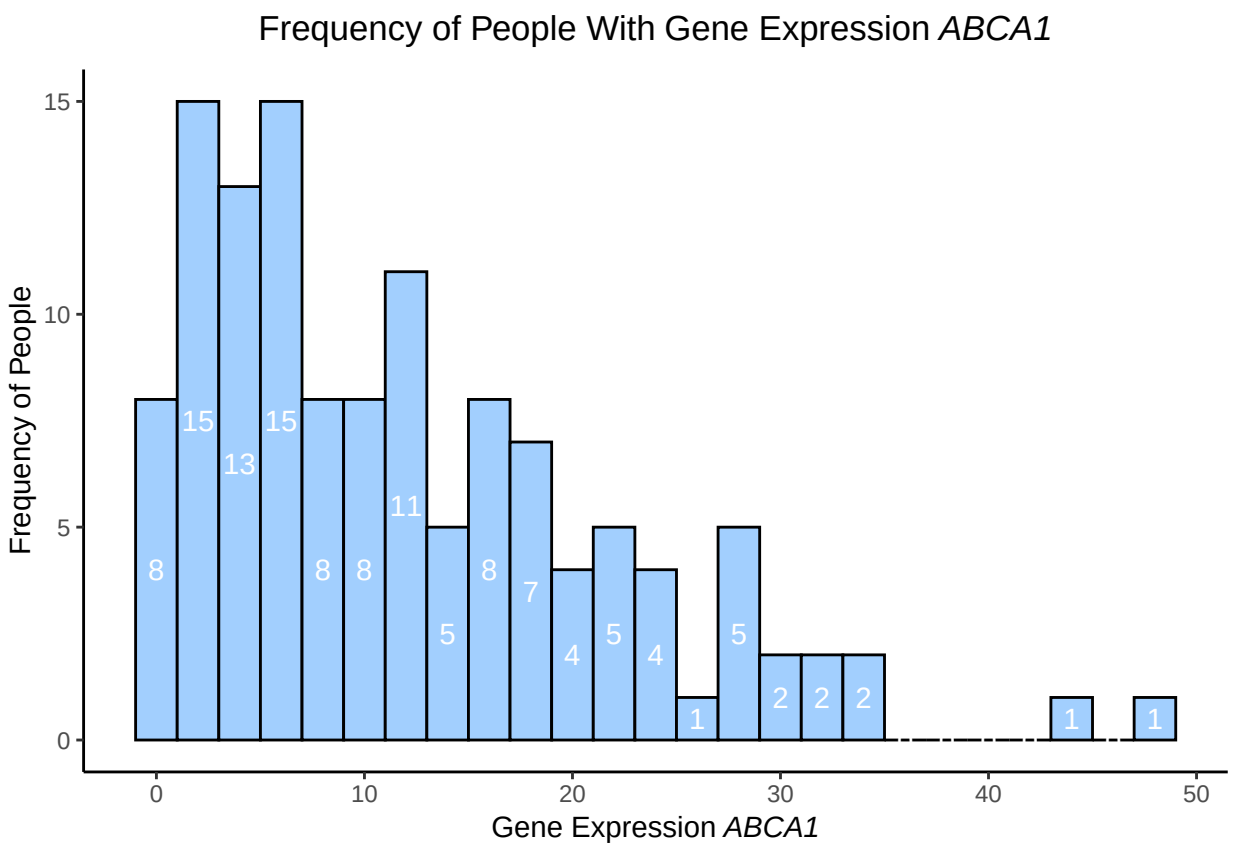
```r
#define color palette
my_palette1 <- c("#A2CFFE") # Choose any colors you like
```

```r
#load the required package
library(ggplot2)
```

```r
#create italicize labels
#my_x1 <- expression(paste("Gene Expression ", italic("ABCA1")))
#my_title1 <- expression(paste("Frequency of People With Gene Expression ", italic("ABCA1")))
```

```r
#creating function to make histogram for gene expression

create_histogram <- function(df, gene){
  italic.gene <- c(gene)
  regular.lab <- c("Frequency of People With", "Gene Expression")
  my_title1 <- eval(bquote(expression(.(regular.lab[1]) ~.(regular.lab[2]) ~italic(.(italic.gene[1])))))
  my_x1 <- eval(bquote(expression(.(regular.lab[2]) ~italic(.(italic.gene[1])))))

  my_histogram <- ggplot(df, aes(x=.data[[gene]])) +
  geom_histogram(binwidth = 2, fill = my_palette1, col="black") +
  stat_bin(binwidth=2, geom='text', color='white', aes(label=..count..),
           position=position_stack(vjust = 0.5)) +
  labs(x = my_x1,y = 'Frequency of People', title = my_title1) +
  theme(plot.title = element_text(hjust = 0.5)) +
  newBlankTheme

  my_histogram

}

create_histogram(clean_data, "ABCA1")
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
#define color palette
my_palette2 <- c("#D697C1") # Choose any colors you like
```

```
#create italicize labels
#my_y2 <- expression(paste("Gene Expression ", italic("ABCA1")))
#my_title2 <- expression(paste("Gene Expression ", italic("ABCA1"), " By Age"))
```

```
#creating function to make scatterplot for gene expression

create_scatterplot <- function(df, gene, cont.var){
  italic.gene <- c(gene)
  regular.lab <- c("Scatterplot of", "Gene Expression", "By Age")
  my_title2 <- eval(bquote(expression(.(regular.lab[1]) ~.(regular.lab[2]) ~italic(.(italic.gene[1])) ~
  my_y2 <- eval(bquote(expression(.(regular.lab[2]) ~italic(.(italic.gene[1]))))))

  my_scatterplot <- ggplot(df,aes(x=.data[[cont.var]],y=.data[[gene]])) +
  geom_point(color=my_palette2) +
  labs(x = "Age", y = my_y2, title = my_title2) +
  theme(plot.title = element_text(hjust = 0.5)) +
  newBlankTheme

  my_scatterplot

}

create_scatterplot(clean_data, "ABCA1", cont.var = "age")
```

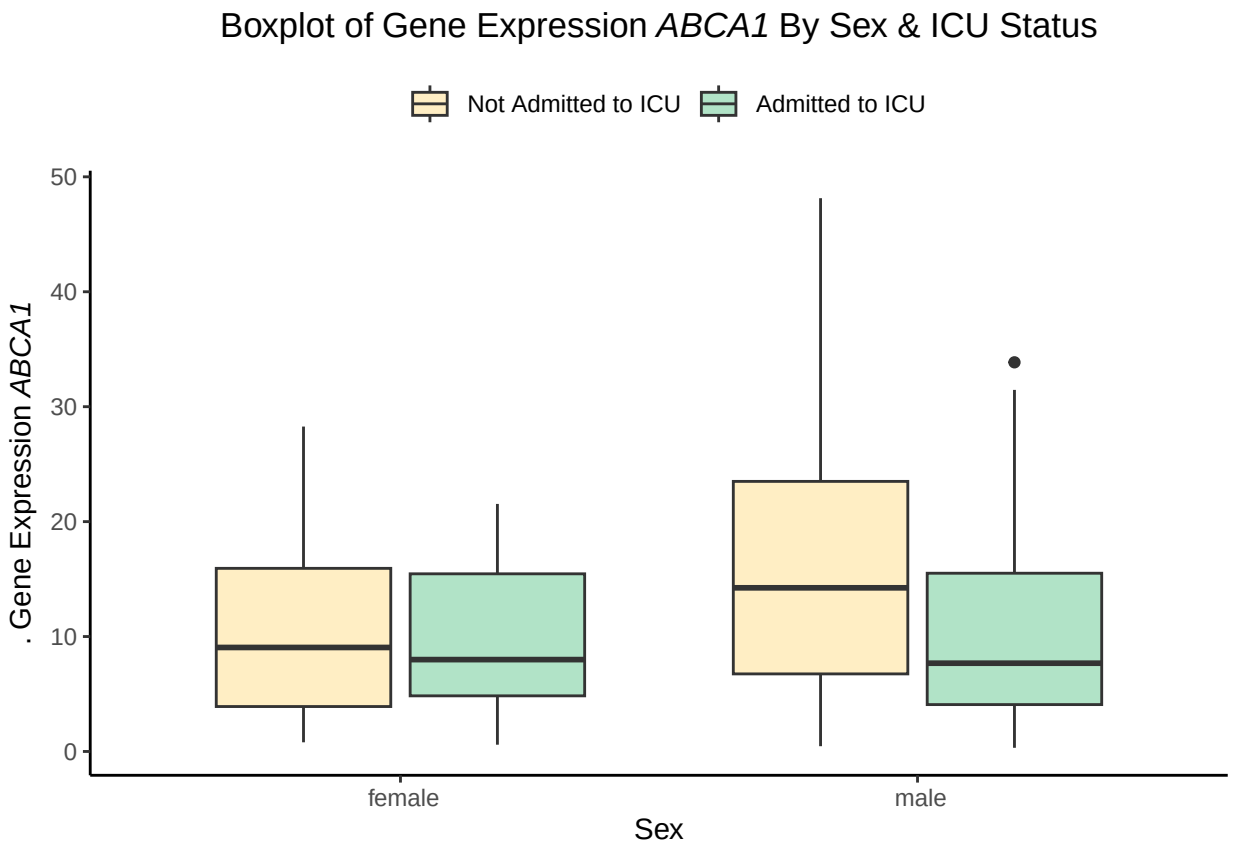
```
## Warning: Removed 3 rows containing missing values (`geom_point()`).
```


```
#define color palette
my_palette3 <- c("#FFEEC4","#B1E3C7") # Choose any colors you like
```

```
#create italicize labels
#my_y3 <- expression(paste("Gene Expression ", italic("ABCA1")))
#my_title3 <- expression(paste("Gene Expression ", italic("ABCA1"), " By Sex & ICU Status"))
```

```
#creating function to make boxplot for gene expression

create_boxplot <- function(df, gene, x.cat, color.cat){
  italic.gene <- c(gene)
  regular.lab <- c("Boxplot of", "Gene Expression", "By Sex & ICU Status")
  my_title3 <- eval(bquote(expression(.(regular.lab[1]) ~.(regular.lab[2]) ~italic(.(italic.gene[1])) ~
  my_y3 <- eval(bquote(expression(. ~.(regular.lab[2]) ~italic(.(italic.gene[1]))))))


  my_boxplot <- ggplot(df, aes(x=.data[[x.cat]], y=.data[[gene]], fill=.data[[color.cat]])) +
  theme(legend.title = element_blank()) +
  geom_boxplot() +
  scale_fill_manual(values = my_palette3, labels = c("Not Admitted to ICU", "Admitted to ICU")) +
  labs(x = "Sex",
       y = my_y3,
       fill = "ICU Status",
       title = my_title3) +
```

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

  my_boxplot

}

create_boxplot(clean_data, "ABCA1", x.cat = "sex", color.cat = "icu_status")
```



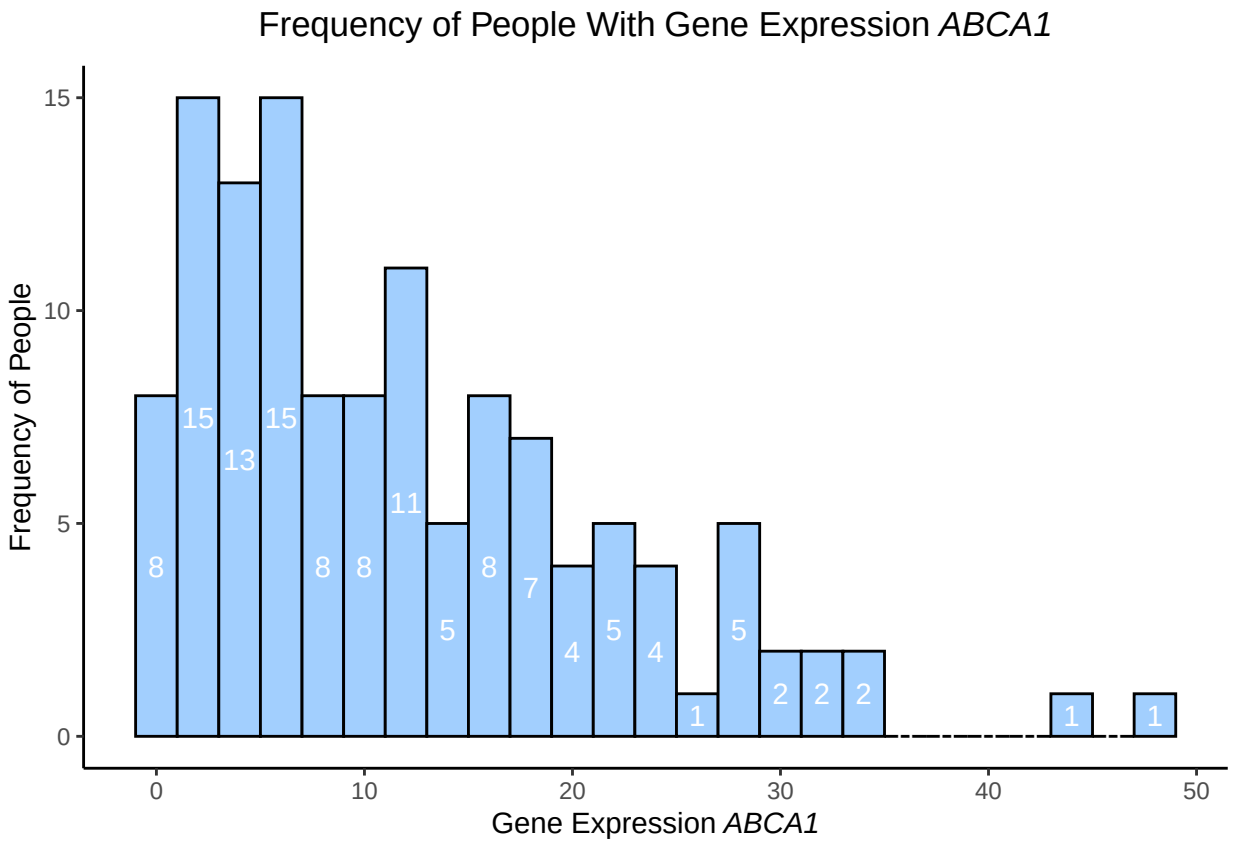

```
#this function takes in the name of the df, list of one or more gene names, one continuous covariate, a
all_plots <- function(df, specific.g, cont.var1, cat.var2){
  h_plot <- create_histogram(df, specific.g)
  s_plot <- create_scatterplot(df, specific.g, cont.var1)
  b_plot <- create_boxplot(df, specific.g, x.cat = cat.var2[1], color.cat = cat.var2[2])

  plot_list <- list(h_plot, s_plot, b_plot)
  plot_list
}
```

```
spec.cat <- c("sex", "icu_status")
spec.cont <- c("age")
spec.gene <- c("ABCA1", "ABCA2", "ABCA3")
```

```
#implement a loop through all of the selected genes to generate my figures using the function I created
for (i in 1:length(spec.gene)){
  plot_gene <- all_plots(clean_data, spec.gene[i], spec.cont, spec.cat)
```

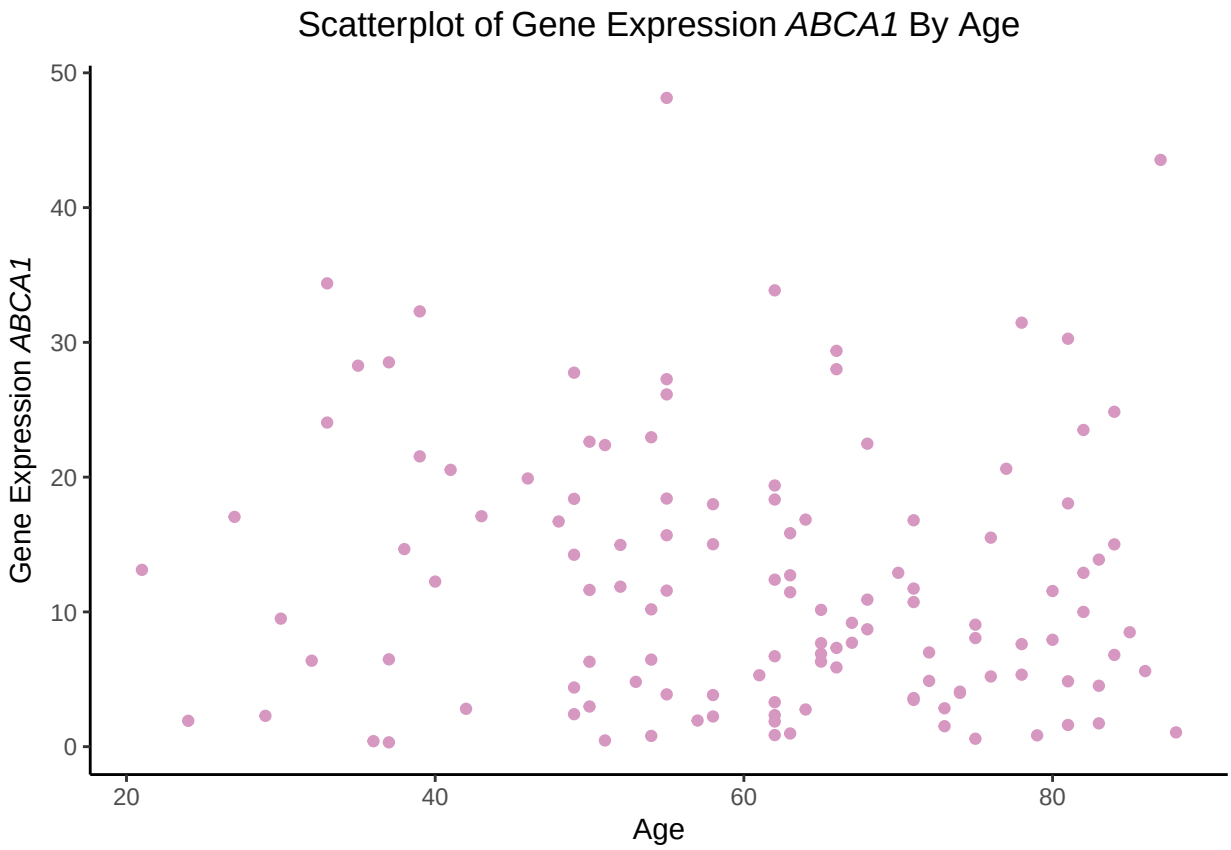
```
  print(plot_gene)
}
```

```
## [[1]]
```

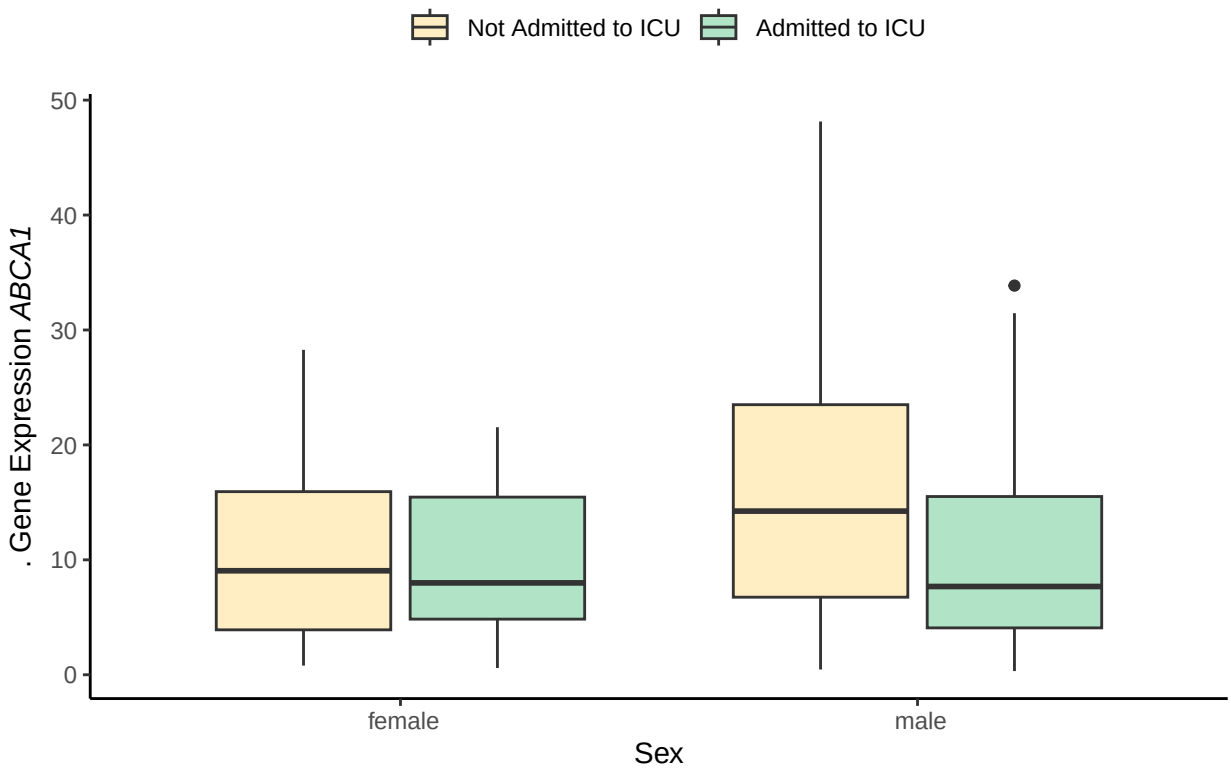



```
##
## [[2]]
```

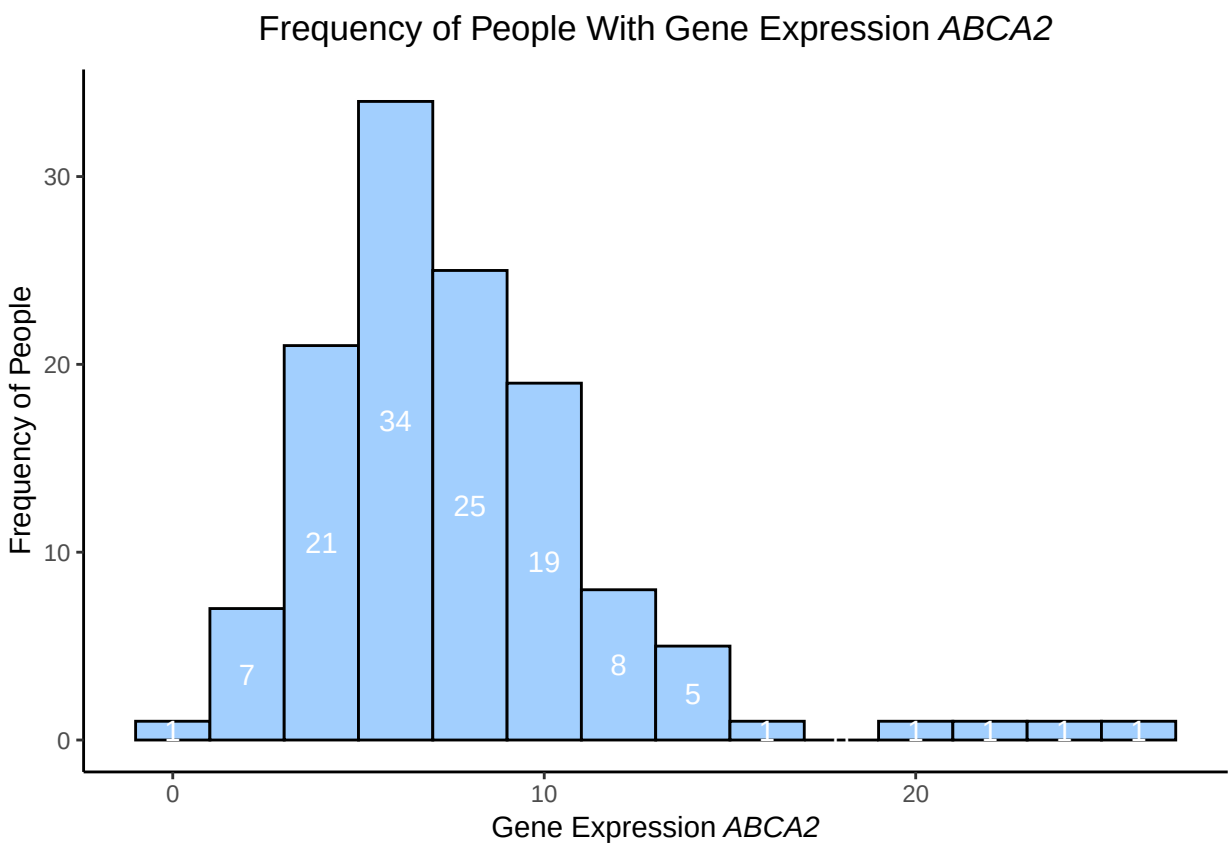
```
## Warning: Removed 3 rows containing missing values (‘geom_point()‘).
```


Scatterplot of Gene Expression ABCA1 By Age
Gene Expression ABCA1
Age
50
40
30
20
10
0
20
40
60
80

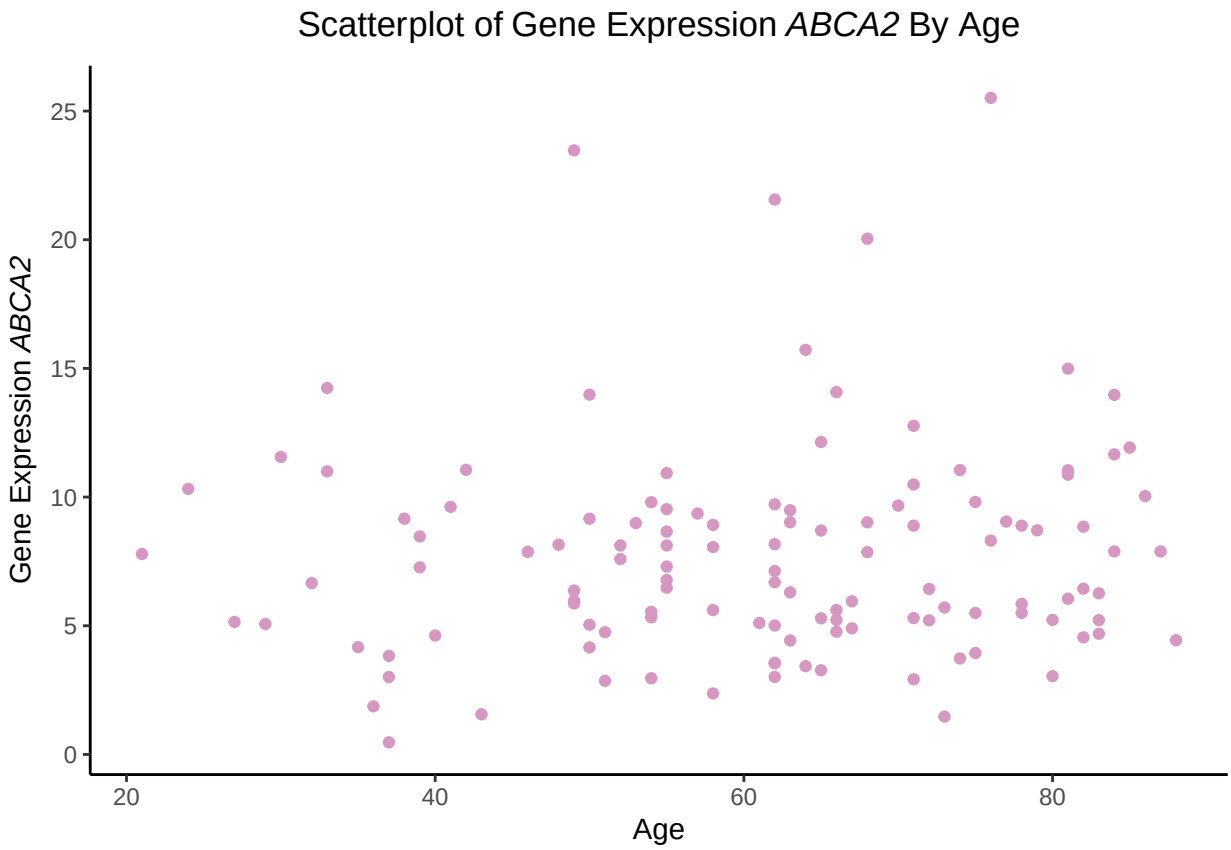
```
##
## [[3]]
```


Boxplot of Gene Expression ABCA1 By Sex & ICU Status
Not Admitted to ICU
Admitted to ICU
50
40
30
20
10
0
. Gene Expression ABCA1
female
male
Sex

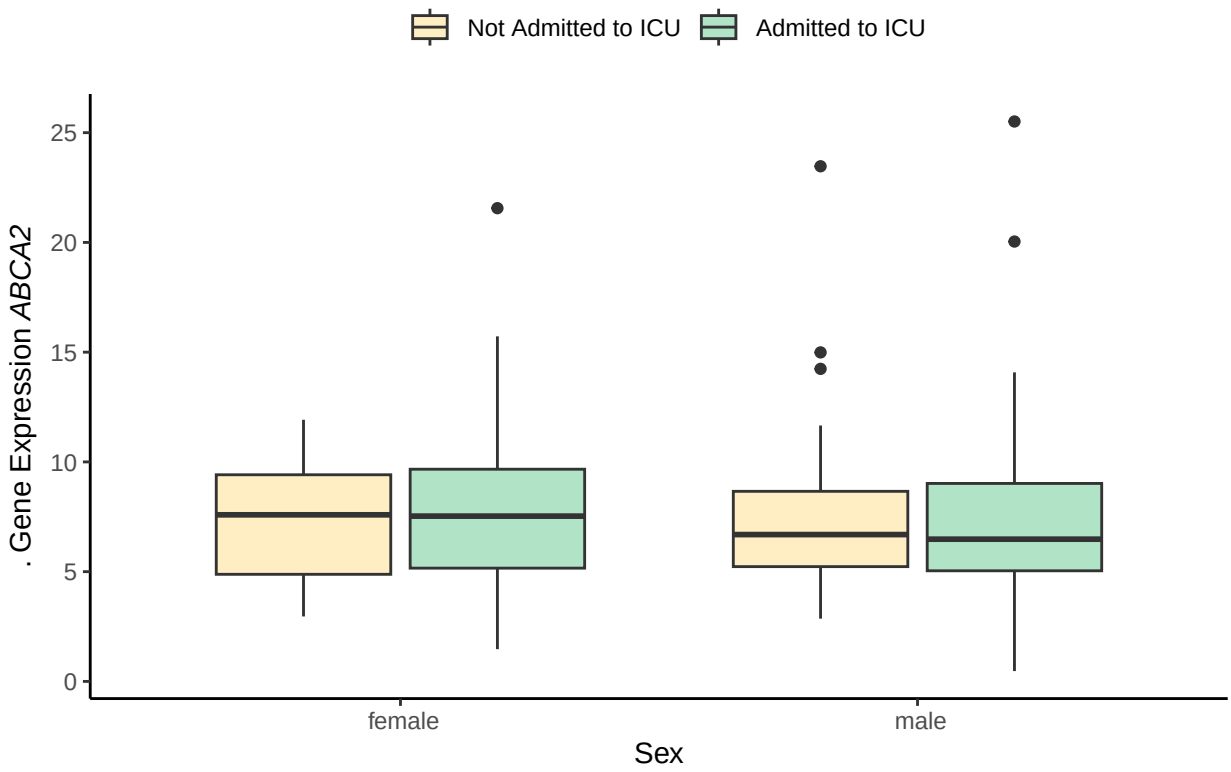
```
##
## [[1]]
```



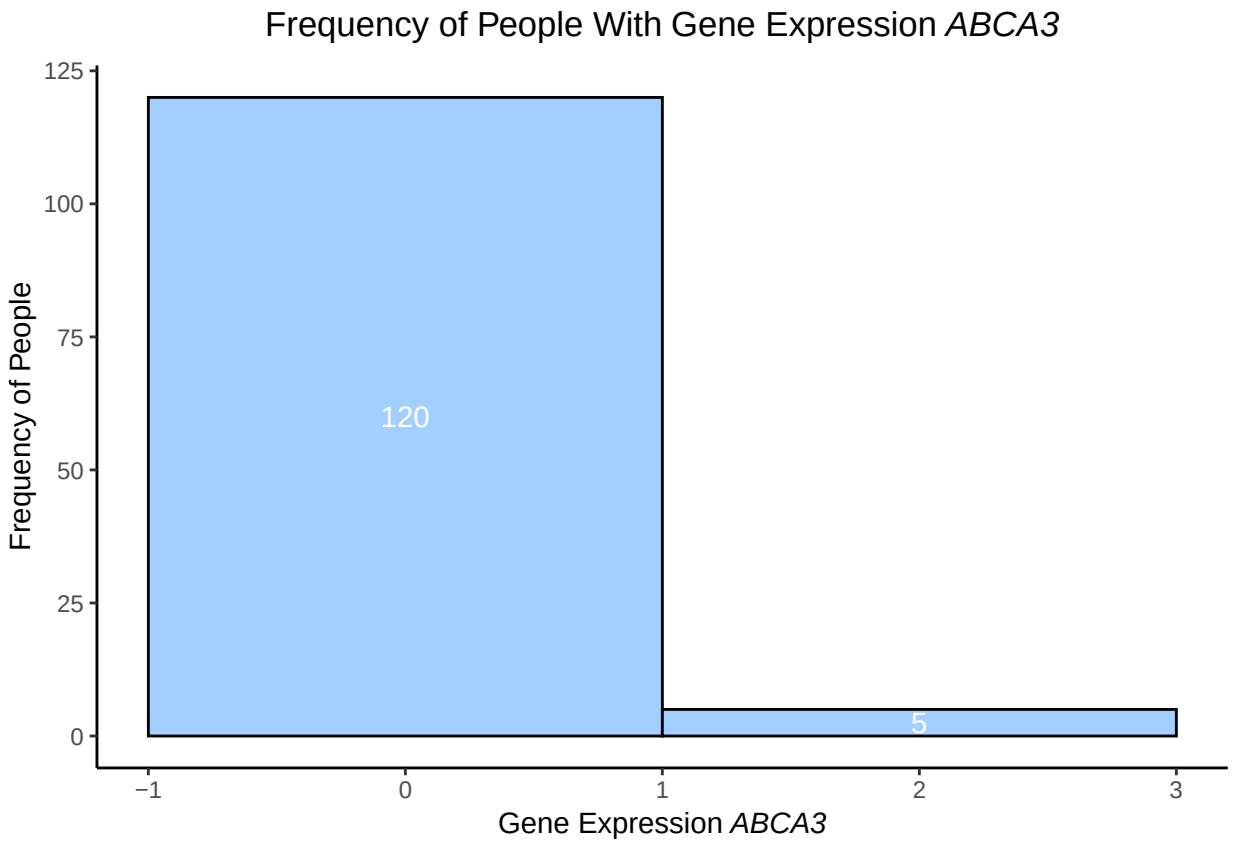

```
##
## [[2]]

## Warning: Removed 3 rows containing missing values (‘geom_point()‘).
```


Scatterplot of Gene Expression ABCA2 By Age
Gene Expression ABCA2
25
20
15
10
5
0
20
40
60
80
Age

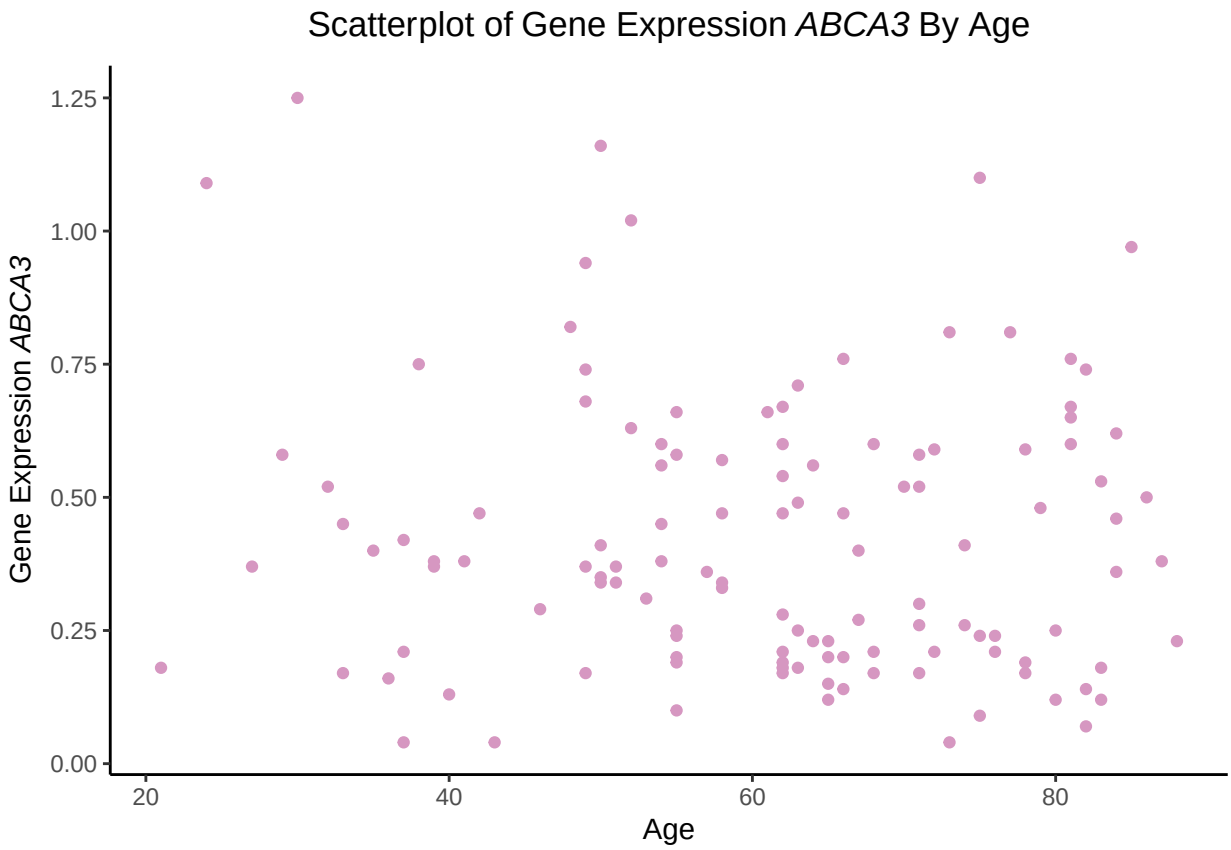
```
##
## [[3]]
```


Boxplot of Gene Expression ABCA2 By Sex & ICU Status
Not Admitted to ICU
Admitted to ICU
. Gene Expression ABCA2
25
20
15
10
5
0
female
male
Sex

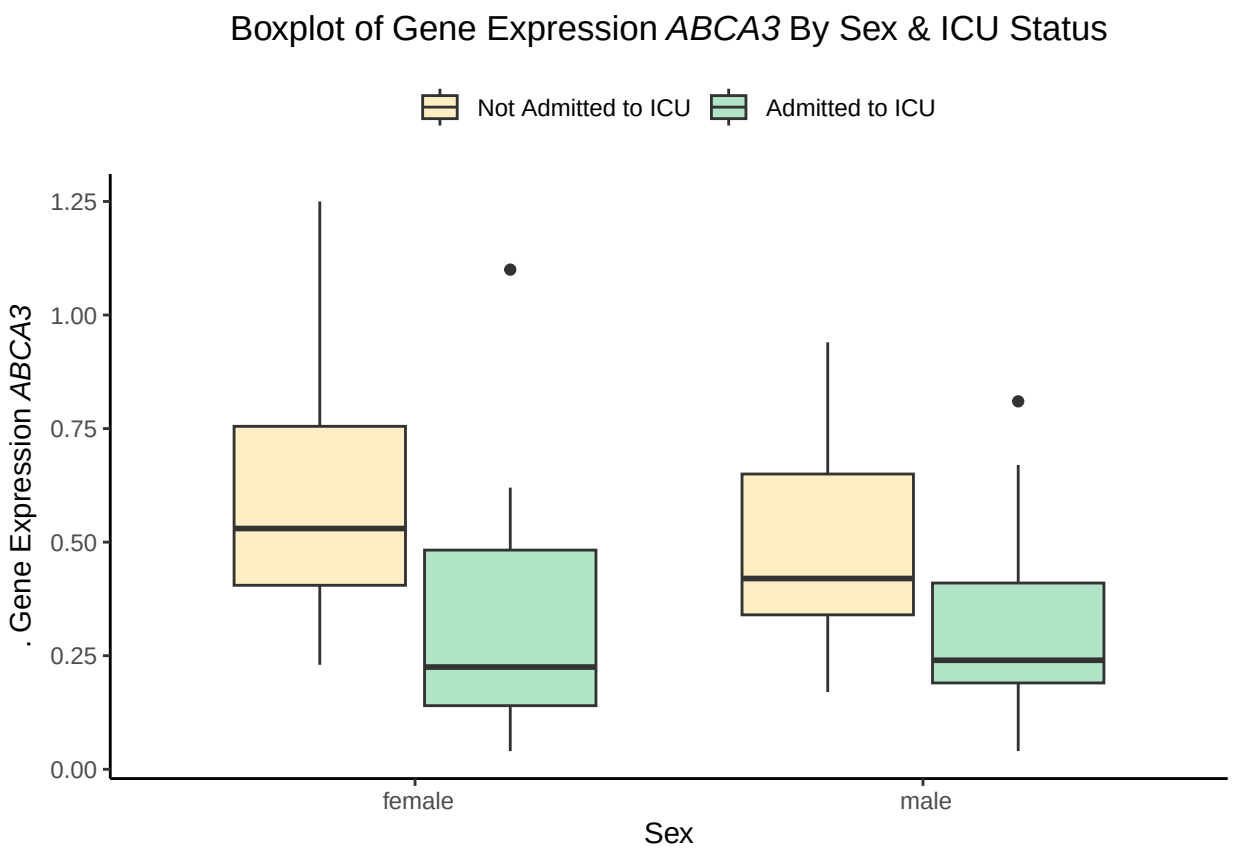
```
##
## [[1]]
```




```
##
## [[2]]

## Warning: Removed 3 rows containing missing values (‘geom_point()‘).
```


Scatterplot of Gene Expression ABCA3 By Age
Gene Expression ABCA3
1.25
1.00
0.75
0.50
0.25
0.00
20
40
60
80
Age

```
##
## [[3]]
```


Boxplot of Gene Expression ABCA3 By Sex & ICU Status
Not Admitted to ICU
Admitted to ICU
. Gene Expression ABCA3
1.25
1.00
0.75
0.50
0.25
0.00
female
male
Sex