# Project Presentation

Sai Lakkireddy

2023-08-08

## Chosen Gene: AAMP

**AAMP stands for "Angio-Associated Migratory Cell Protein" gene. This gene is responsible for making the AAMP protein, which is involved in cell movement and angiogenesis, contributing to processes like wound healing and tissue repair. Genes and their corresponding proteins are crucial for the proper functioning of our bodies.**

**In this analysis, we study the association between AAMP's gene expression, age, COVID Status and ICU Status**

## Importing and combining the Data from two csv files

Steps:
1. Import both the csv files
2. Convert the gene expression into long format
3. Inner join it with meta data

```
#set current working directory to the previous folder
setwd("../")

#import the geneExpression and metaData csv from the data folder
data_gene_Expression <- read.csv("data/QBS103_finalProject_geneExpression.csv", header=TRUE)
data_meta <- read.csv("data/QBS103_finalProject_metadata.csv", header=TRUE)


#We convert the geneExpression data from wide form into long form
data_gene_Expression.longForm <- data_gene_Expression %>%
  pivot_longer(cols = starts_with(c("COVID_","NONCOVID_")),
  names_to = "participant_id",
  values_to = "gene_expression_value"
  )

#make a final data frame by combining the two data sets and making it a data frame
final_df <- as.data.frame(data_gene_Expression.longForm %>% inner_join( data_meta,
           by=c('participant_id')))

head(final_df)
```

```
##       X             participant_id gene_expression_value geo_accession
## 1 APOA1   COVID_01_39y_male_NonICU                  0.00    GSM4753021
## 2 APOA1   COVID_02_63y_male_NonICU                  0.12    GSM4753022
## 3 APOA1   COVID_03_33y_male_NonICU                  0.00    GSM4753023
## 4 APOA1   COVID_04_49y_male_NonICU                  0.09    GSM4753024
## 5 APOA1   COVID_05_49y_male_NonICU                  0.08    GSM4753025
## 6 APOA1 COVID_07_38y_female_NonICU                  0.00    GSM4753027
##                  status X.Sample_submission_date last_update_date type
## 1 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 2 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 3 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 4 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 5 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 6 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
##   channel_count             source_name_ch1 organism_ch1
## 1             1 Leukocytes from whole blood Homo sapiens
## 2             1 Leukocytes from whole blood Homo sapiens
## 3             1 Leukocytes from whole blood Homo sapiens
## 4             1 Leukocytes from whole blood Homo sapiens
## 5             1 Leukocytes from whole blood Homo sapiens
## 6             1 Leukocytes from whole blood Homo sapiens
##            disease_status age     sex icu_status apacheii charlson_score
## 1 disease state: COVID-19  39    male         no       15              0
## 2 disease state: COVID-19  63    male         no  unknown              2
## 3 disease state: COVID-19  33    male         no  unknown              2
## 4 disease state: COVID-19  49    male         no  unknown              1
## 5 disease state: COVID-19  49    male         no       19              1
## 6 disease state: COVID-19  38  female         no  unknown              7
##   mechanical_ventilation ventilator.free_days
## 1                    yes                    0
## 2                     no                   28
## 3                     no                   28
## 4                     no                   28
## 5                    yes                   23
## 6                     no                   28
##   hospital.free_days_post_45_day_followup ferritin.ng.ml. crp.mg.l.
## 1                                      0             946      73.1
## 2                                     39            1060   unknown
## 3                                     18            1335      53.2
## 4                                     39             583     251.1
## 5                                     27             800     355.8
## 6                                     42             366   unknown
##   ddimer.mg.l_feu. procalcitonin.ng.ml.. lactate.mmol.l. fibrinogen     sofa
## 1              1.3                    36            0.9        513        8
## 2             1.03                  0.37         unknown    unknown  unknown
## 3             1.48                  0.07         unknown        513  unknown
## 4             1.32                  0.98            0.87        949  unknown
## 5             0.69                  4.92            1.48        929        7
## 6             0.87                  0.06            1.17        478  unknown
```

## Pre-processing the data

Steps:
1. Remove "unknown" strings and prefixes
2. Convert the class the columns to their appropriate type

```
#rename with x column with gene
final_df <- rename(final_df, gene = X)

#remove all unknown strings and substitute it with NAs
final_df[, 16:27][final_df[, 16:27] == ' unknown' | final_df[, 16:27] == 'unknown'] <- NA

#format the disease status column to just include the status
final_df$disease_status <- sub('disease state: ', '', final_df$disease_status)

#convert the column type of disease_status, sex, icu_status and mechanical_ventilation to factor
final_df <- final_df %>%
  mutate_at(vars(disease_status, sex, icu_status, mechanical_ventilation), as.factor)

#convert the class of age, charlson_score
final_df <- final_df %>%
  mutate_at(vars(age, apacheii,ferritin.ng.ml.,
                 crp.mg.l., ddimer.mg.l_feu.,
                 procalcitonin.ng.ml.., lactate.mmol.l., fibrinogen, sofa), as.integer)

head(final_df)
```

```
##    gene             participant_id gene_expression_value geo_accession
## 1 APOA1   COVID_01_39y_male_NonICU                 0.00    GSM4753021
## 2 APOA1   COVID_02_63y_male_NonICU                 0.12    GSM4753022
## 3 APOA1   COVID_03_33y_male_NonICU                 0.00    GSM4753023
## 4 APOA1   COVID_04_49y_male_NonICU                 0.09    GSM4753024
## 5 APOA1   COVID_05_49y_male_NonICU                 0.08    GSM4753025
## 6 APOA1 COVID_07_38y_female_NonICU                 0.00    GSM4753027
##                  status X.Sample_submission_date last_update_date type
## 1 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 2 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 3 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 4 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 5 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 6 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
##   channel_count             source_name_ch1 organism_ch1 disease_status age
## 1             1 Leukocytes from whole blood Homo sapiens       COVID-19  39
## 2             1 Leukocytes from whole blood Homo sapiens       COVID-19  63
## 3             1 Leukocytes from whole blood Homo sapiens       COVID-19  33
## 4             1 Leukocytes from whole blood Homo sapiens       COVID-19  49
## 5             1 Leukocytes from whole blood Homo sapiens       COVID-19  49
## 6             1 Leukocytes from whole blood Homo sapiens       COVID-19  38
##      sex icu_status apacheii charlson_score mechanical_ventilation
## 1   male         no       15              0                   yes
## 2   male         no       NA              2                    no
## 3   male         no       NA              2                    no
## 4   male         no       NA              1                    no
```

```
## 5    male         no       19              1                    yes
## 6  female         no       NA              7                     no
##   ventilator.free_days hospital.free_days_post_45_day_followup ferritin.ng.ml.
## 1                    0                                       0            946
## 2                   28                                      39           1060
## 3                   28                                      18           1335
## 4                   28                                      39            583
## 5                   23                                      27            800
## 6                   28                                      42            366
##   crp.mg.l. ddimer.mg.l_feu. procalcitonin.ng.ml.. lactate.mmol.l. fibrinogen
## 1        73               1                    36              0        513
## 2        NA               1                     0             NA         NA
## 3        53               1                     0             NA        513
## 4       251               1                     0              0        949
## 5       355               0                     4              1        929
## 6        NA               0                     0              1        478
##   sofa
## 1    8
## 2   NA
## 3   NA
## 4   NA
## 5    7
## 6   NA
```

## Optional - handle missing values

```r
# check all the numeric colums
num_cols <- names(select_if(final_df, is.numeric))

# Create an imputation model
imputation_model <- mice(final_df[num_cols], method = "pmm", printFlag = FALSE)

# Perform the imputation
imputed_data_final <- complete(imputation_model)
# update the final data frame with the imputed values
final_df[num_cols] <- imputed_data_final[num_cols]

head(final_df)
```

```
##    gene             participant_id gene_expression_value geo_accession
## 1 APOA1   COVID_01_39y_male_NonICU                  0.00    GSM4753021
## 2 APOA1   COVID_02_63y_male_NonICU                  0.12    GSM4753022
## 3 APOA1   COVID_03_33y_male_NonICU                  0.00    GSM4753023
## 4 APOA1   COVID_04_49y_male_NonICU                  0.09    GSM4753024
## 5 APOA1   COVID_05_49y_male_NonICU                  0.08    GSM4753025
## 6 APOA1 COVID_07_38y_female_NonICU                  0.00    GSM4753027
##                   status X.Sample_submission_date last_update_date type
## 1 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 2 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 3 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 4 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 5 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
```

```
## 6 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
##   channel_count             source_name_ch1 organism_ch1 disease_status age
## 1             1 Leukocytes from whole blood Homo sapiens       COVID-19  39
## 2             1 Leukocytes from whole blood Homo sapiens       COVID-19  63
## 3             1 Leukocytes from whole blood Homo sapiens       COVID-19  33
## 4             1 Leukocytes from whole blood Homo sapiens       COVID-19  49
## 5             1 Leukocytes from whole blood Homo sapiens       COVID-19  49
## 6             1 Leukocytes from whole blood Homo sapiens       COVID-19  38
##      sex icu_status apacheii charlson_score mechanical_ventilation
## 1   male         no       15              0                   yes
## 2   male         no        4              2                    no
## 3   male         no       12              2                    no
## 4   male         no       12              1                    no
## 5   male         no       19              1                   yes
## 6 female         no       13              7                    no
##   ventilator.free_days hospital.free_days_post_45_day_followup ferritin.ng.ml.
## 1                    0                                       0             946
## 2                   28                                      39            1060
## 3                   28                                      18            1335
## 4                   28                                      39             583
## 5                   23                                      27             800
## 6                   28                                      42             366
##   crp.mg.l. ddimer.mg.l_feu. procalcitonin.ng.ml.. lactate.mmol.l. fibrinogen
## 1        73               1                    36              0        513
## 2       154               1                     0              1        458
## 3        53               1                     0              0        513
## 4       251               1                     0              0        949
## 5       355               0                     4              1        929
## 6        19               0                     0              1        478
##   sofa
## 1    8
## 2    1
## 3   10
## 4   10
## 5    7
## 6   10
```

### Create a subset the AAMP Gene and the chosen covariates

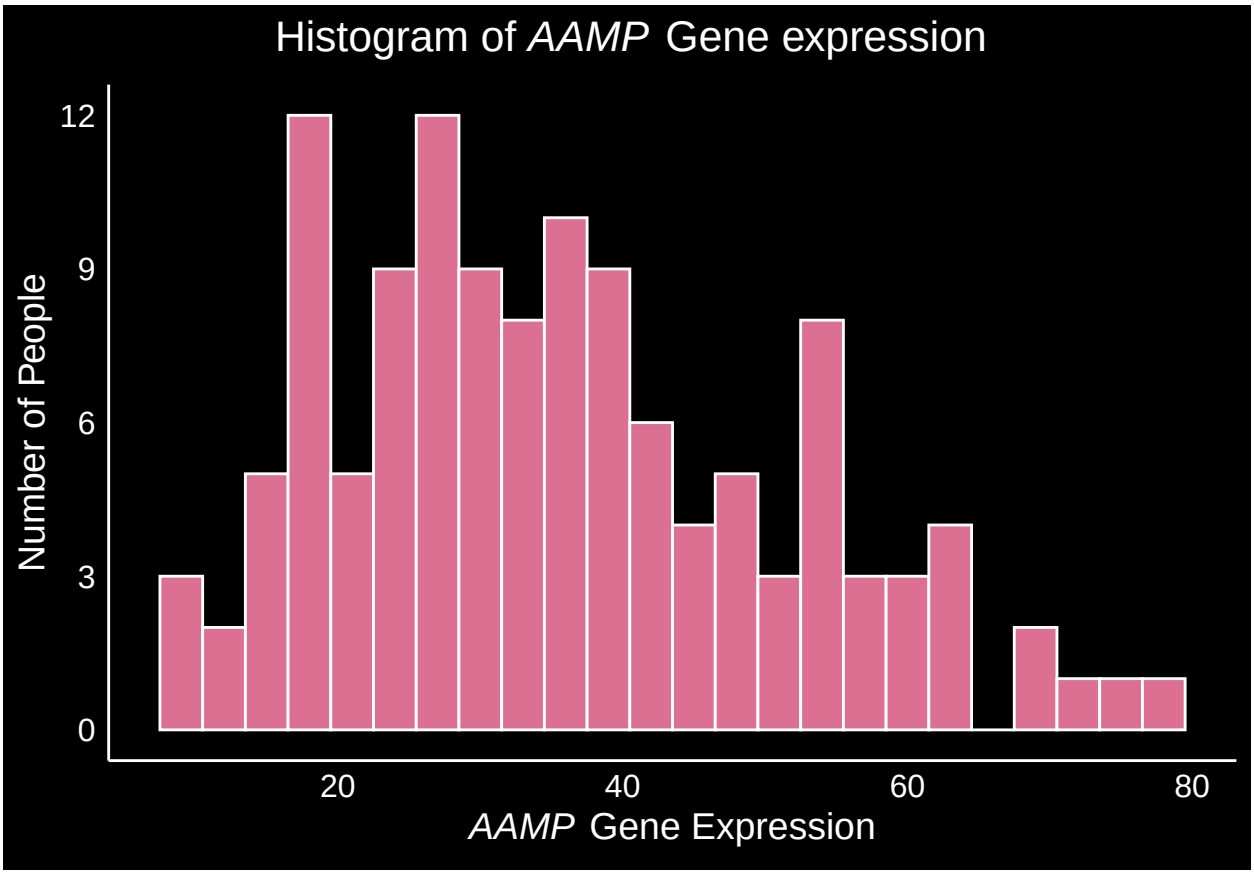
```
final_subset <- final_df[final_df$gene == 'AAMP',
                         c('gene',
                           'gene_expression_value','age',
                           'icu_status', 'disease_status')]

head(final_subset)
```

```
##      gene gene_expression_value age icu_status disease_status
## 4501 AAMP                61.08  39         no       COVID-19
## 4502 AAMP                54.54  63         no       COVID-19
## 4503 AAMP                25.19  33         no       COVID-19
## 4504 AAMP                67.95  49         no       COVID-19
## 4505 AAMP                18.29  49         no       COVID-19
```

```
## 4506 AAMP                51.35  38         no       COVID-19
```

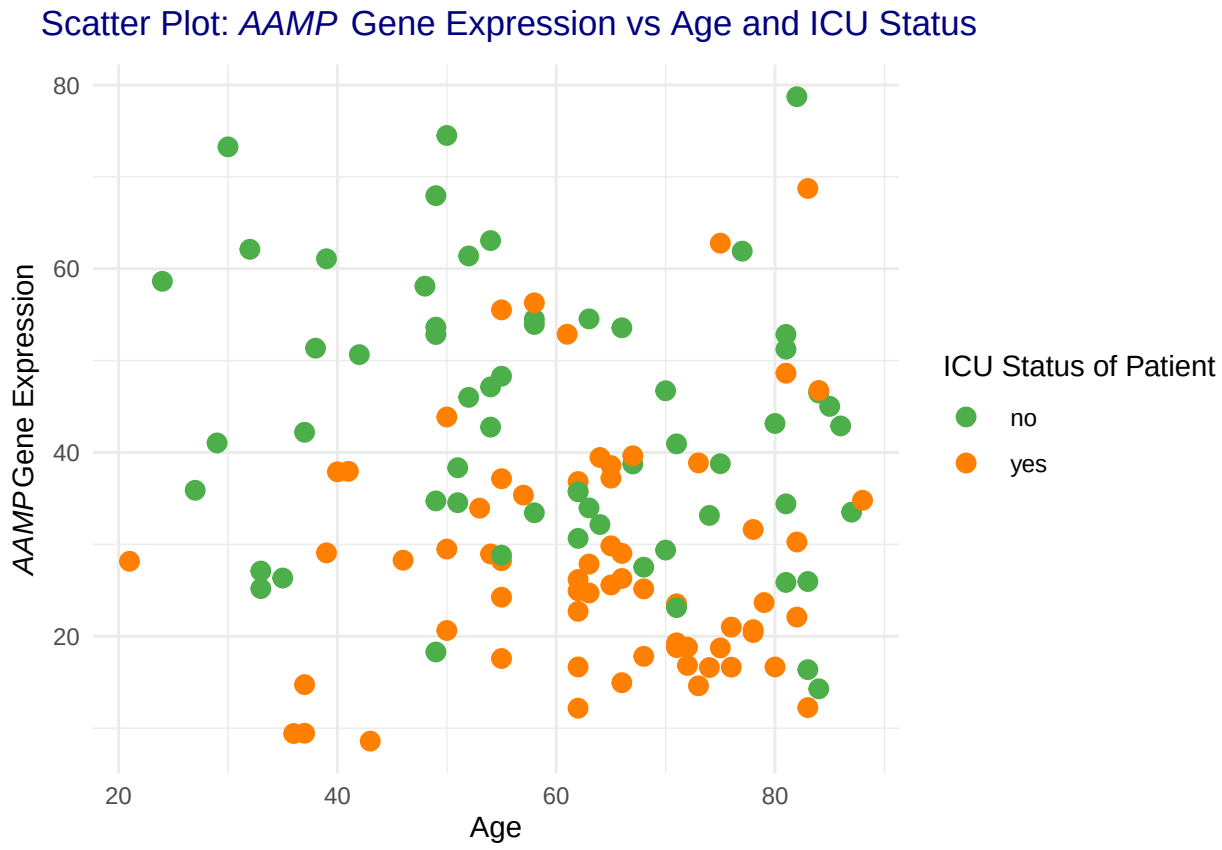
### Histogram for Gene Expression

```
breaks <- seq(0, 15, by = 3)

# Create the histogram with integer bins
ggplot(final_subset, aes(x = gene_expression_value)) +
  geom_histogram(binwidth = 3, color = "white", fill = "#DB7093") +
  scale_y_continuous(breaks = breaks) +
  theme_minimal() +
  theme(
    panel.background = element_rect(fill = "black"),
    plot.background = element_rect(fill = "black"),
    axis.line = element_line(color = "white"),
    axis.text = element_text(color = "white", size = 12),
    axis.title = element_text(color = "white", size = 14),
    panel.grid = element_blank(),
    plot.title = element_text(color = "white", size = 16, face = "bold", hjust = 0.4),
  ) +
  ggtitle(expression(paste("Histogram of ",italic("AAMP")," Gene expression"))) +
  xlab(expression(paste(italic("AAMP"), " Gene Expression"))) +
  ylab("Number of People")
```



### Scatter plot: Age vs Gene Expression factoring for ICU status

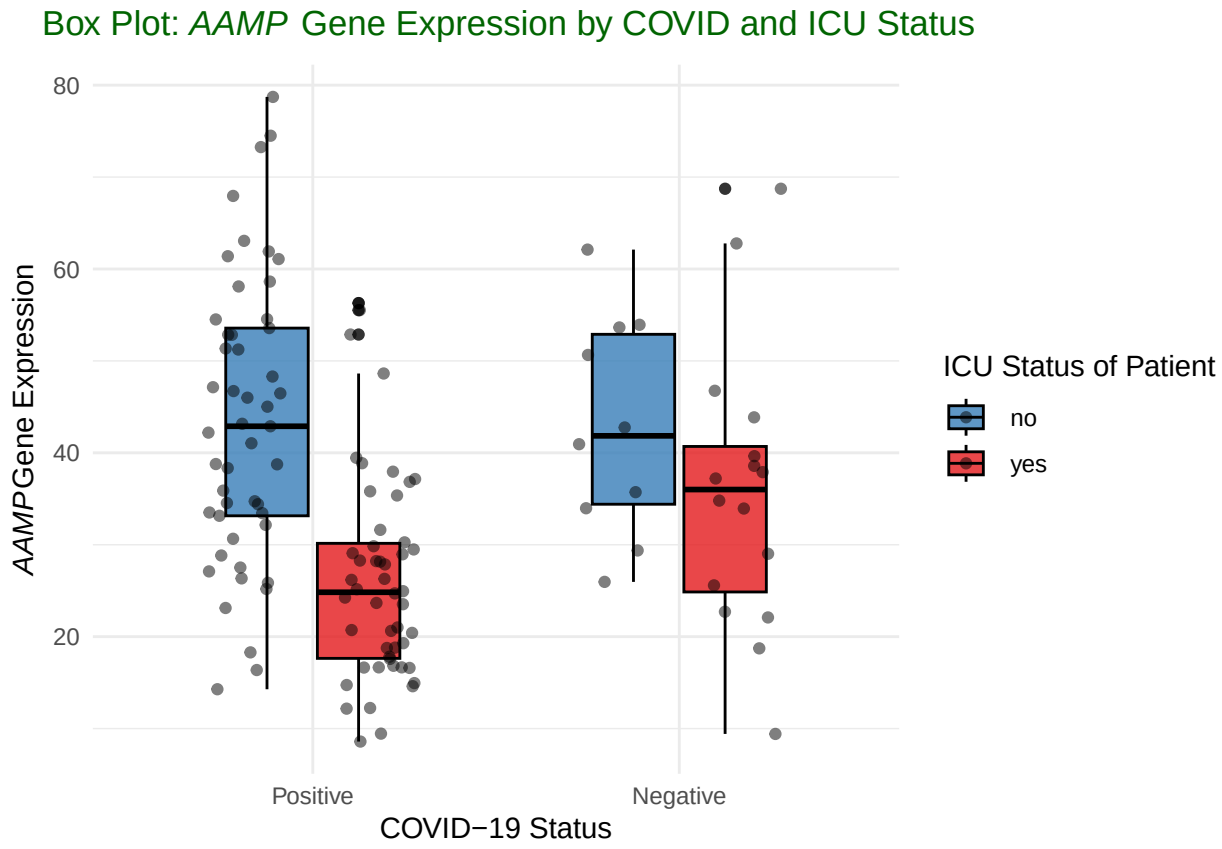
```
my_colors_1 <- c("#4DAF4A", "#FF7F00")


# Create the scatter plot with custom color scheme
ggplot(final_subset, aes(x = age, y = gene_expression_value, color = icu_status)) +
  geom_point(size = 3) +
  scale_color_manual(values = my_colors_1, name = "ICU Status of Patient") +
  theme_minimal() +
      theme(
    plot.title = element_text(color = "navy", size = 13, face = "bold", hjust = 0.4)
  ) +
  ggtitle(expression(paste("Scatter Plot: ",italic("AAMP")," Gene Expression vs Age and ICU Status"))) +
  xlab("Age") +
  ylab(expression(paste(italic("AAMP"),"Gene Expression")))
```




### Box plot: Gene Expression by COVID and ICU Status

```
my_colors_2 <- c("#377eb8", "#e41a1c")
ggplot(final_subset, aes(x = disease_status, y = gene_expression_value, fill = icu_status)) +
  geom_boxplot(color = "black", width = 0.5, alpha = 0.8) +
```

```
  scale_fill_manual(values = my_colors_2, name = "ICU Status of Patient") +
  geom_jitter(position = position_jitterdodge(), alpha = 0.5) +
  scale_x_discrete(labels = c("COVID-19" = "Positive", "non-COVID-19" = "Negative")) +
  theme_minimal() +
    theme_minimal() +
      theme(
    plot.title = element_text(color = "darkgreen", size = 13, face = "bold", hjust = 0.4)
  )+
  ggtitle(expression(paste("Box Plot: ",italic("AAMP")," Gene Expression by COVID and ICU Status"))) +
  xlab("COVID-19 Status") +
  ylab(expression(paste(italic("AAMP"),"Gene Expression")))
```

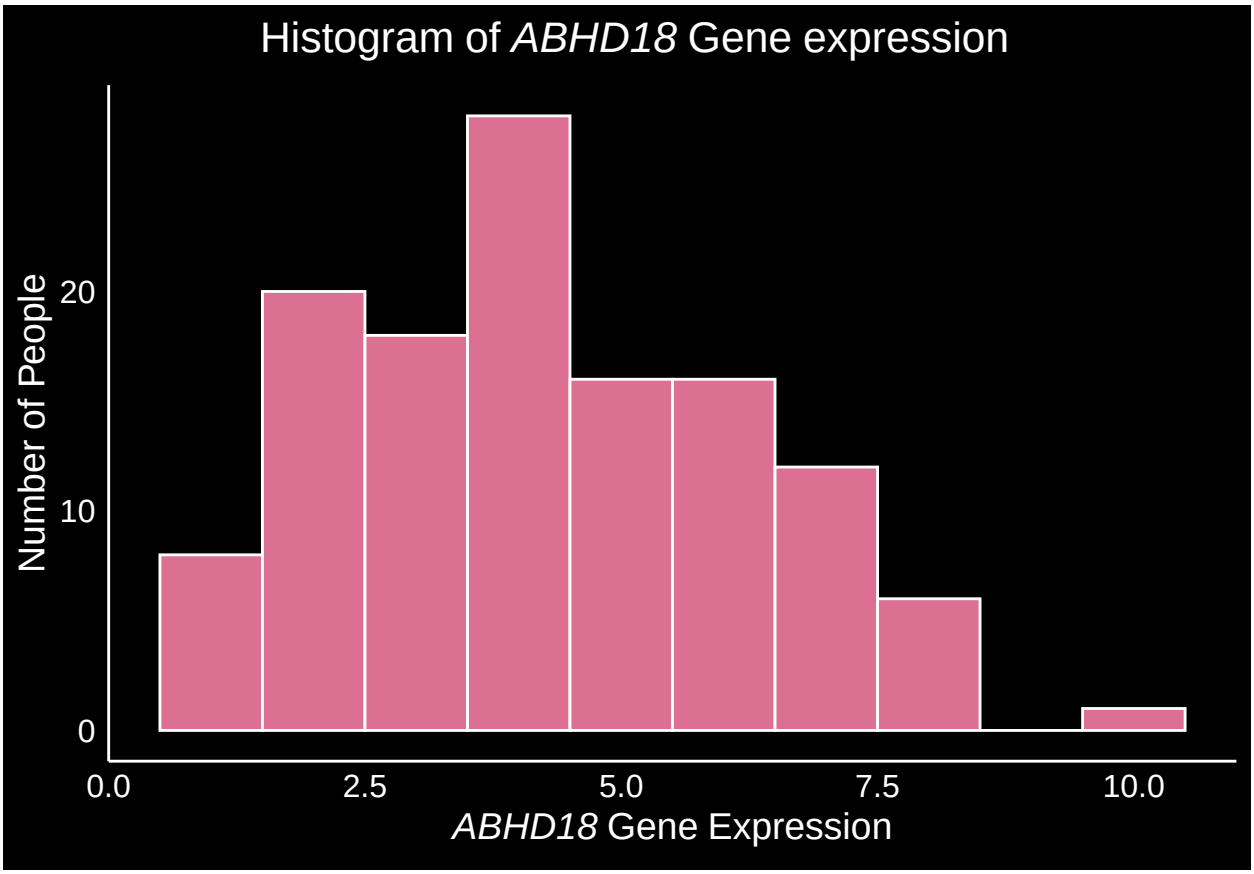


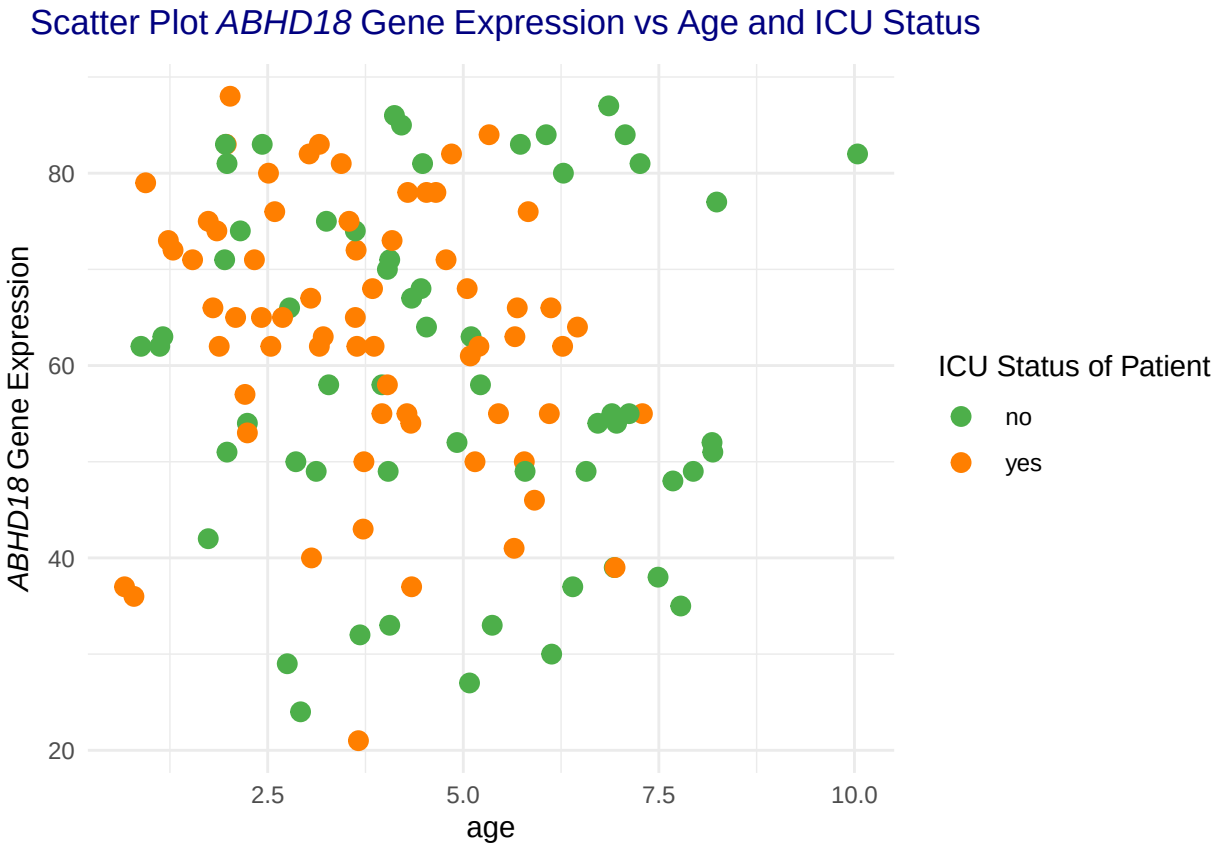

**Plots generated by a function for genes ABHD18, AAMP and ABHD17C**

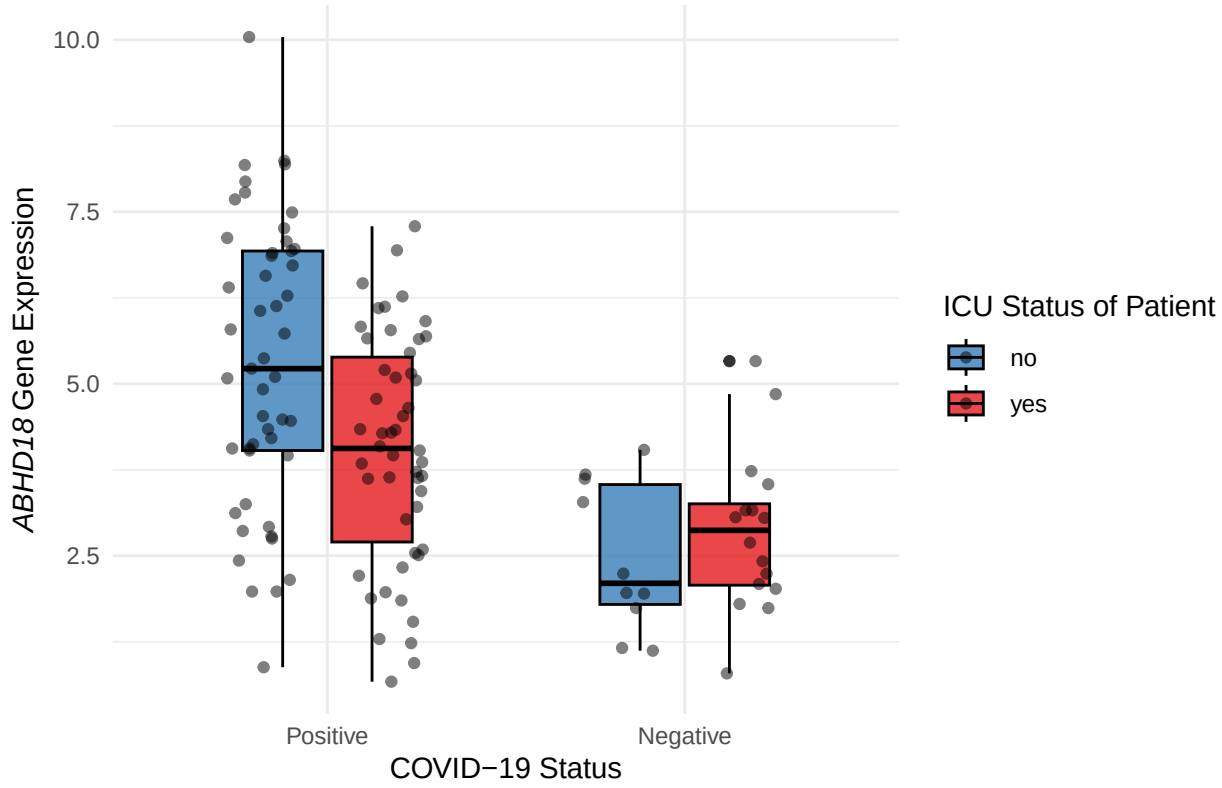
```
my_plots_function <- function(dataFrame, genes.list, cont.covariate, cat.covariates) {
  #breaks <- seq(min(freque), 15, by = 3)
  my_colors_1 <- c("#4DAF4A", "#FF7F00")
  my_colors_2 <- c("#377eb8", "#e41a1c")
  all.plots.list <- list()

  for (gene in genes.list) {
    gene_subset <- final_df[final_df$gene == gene,
                            c('gene','gene_expression_value',paste(cont.covariate), paste(cat.covariates
```

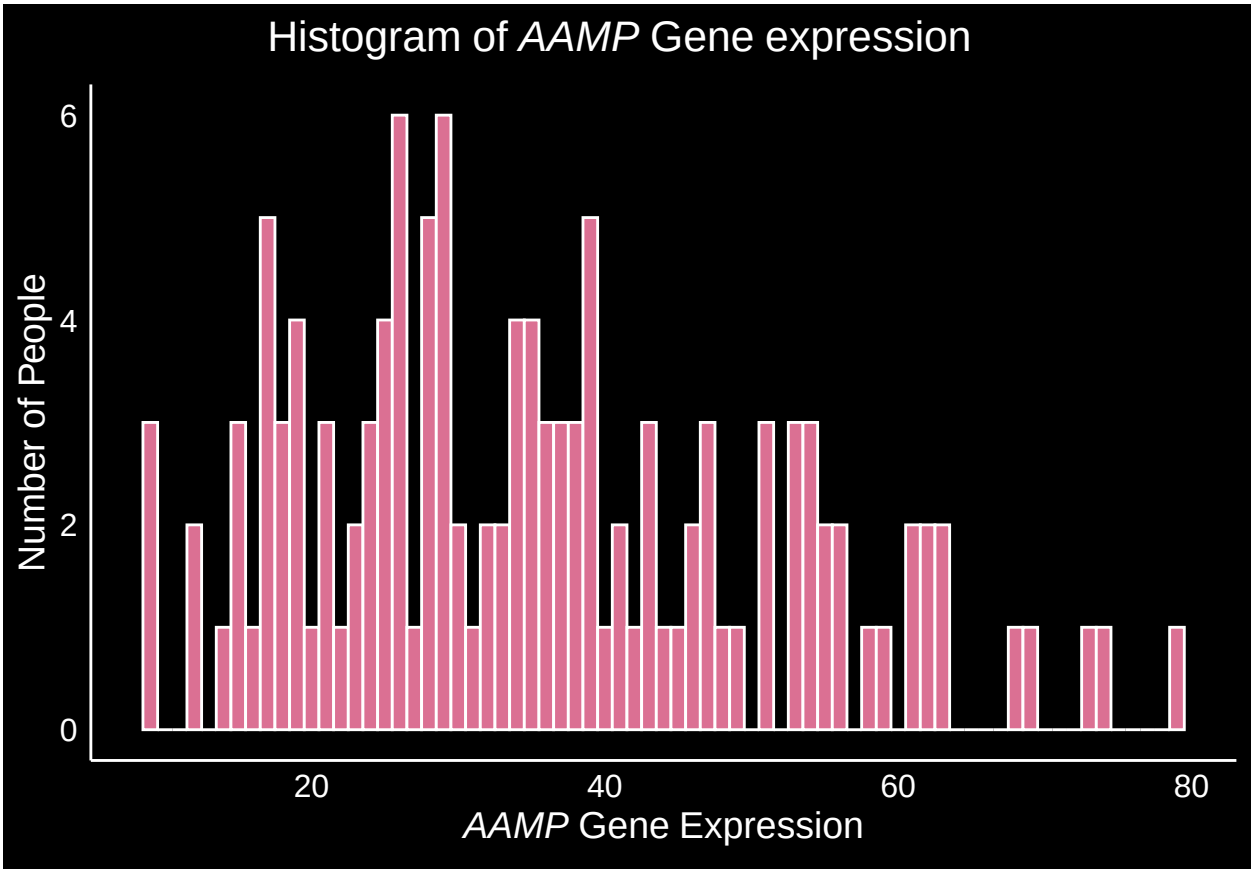
```
   histogram <- ggplot(gene_subset, aes(x = gene_expression_value)) +
  geom_histogram(binwidth = 1, color = "white", fill = "#DB7093") +
  #scale_y_continuous(breaks = breaks) +
  theme_minimal() +
  theme(
    panel.background = element_rect(fill = "black"),
    plot.background = element_rect(fill = "black"),
    axis.line = element_line(color = "white"),
    axis.text = element_text(color = "white", size = 12),
    axis.title = element_text(color = "white", size = 14),
    panel.grid = element_blank(),
    plot.title = element_text(color = "white", size = 16, face = "bold", hjust = 0.4),
  ) +
  ggtitle(substitute(Histogram ~ of ~ italic(gene) ~ Gene ~ expression, list(gene = gene))) +
  xlab(substitute(italic(gene) ~ Gene ~ Expression, list(gene = gene))) +
  ylab("Number of People")


  scatter.plot <- ggplot(gene_subset, aes(x = gene_expression_value, y = gene_subset[[cont.covariate]],
  geom_point(size = 3) +
  scale_color_manual(values = my_colors_1, name = "ICU Status of Patient") +
  theme_minimal() +
      theme(
    plot.title = element_text(color = "navy", size = 13, face = "bold", hjust = 0.4)
  ) +
  ggtitle(substitute(Scatter ~ Plot ~ italic(gene) ~ Gene ~ Expression ~ vs ~ Age ~ and ~ ICU ~ Status)
  xlab(substitute(cont.covariate)) +
  ylab(substitute(italic(gene) ~ Gene ~ Expression))


box.plot <- ggplot(gene_subset, aes(x = gene_subset[[cat.covariates[1]]], y = gene_expression_value, fi
  geom_boxplot(color = "black", width = 0.5, alpha = 0.8) +
  scale_fill_manual(values = my_colors_2, name = "ICU Status of Patient") +
  geom_jitter(position = position_jitterdodge(), alpha = 0.5) +
  scale_x_discrete(labels = c("COVID-19" = "Positive", "non-COVID-19" = "Negative")) +
  theme_minimal() +
  theme(
    plot.title = element_text(color = "darkgreen", size = 13, face = "bold", hjust = 0.4)
  ) +
  ggtitle(substitute(Box ~ plot ~ of ~ italic(gene) ~ Gene ~ Expression ~ by ~ COVID ~ and ~ ICU ~ Statu
  xlab("COVID-19 Status") +
  ylab(substitute(italic(gene) ~ Gene ~ Expression))

    all.plots.list[[gene]] <- list(histogram = histogram, scatter.plot = scatter.plot, box.plot = box.pl
  }

  return(all.plots.list)
}

all.plots.list <- my_plots_function(final_subset, c('ABHD18', 'AAMP', 'ABHD17C'), 'age', c('disease_sta

for (gene in names(all.plots.list)) {
```

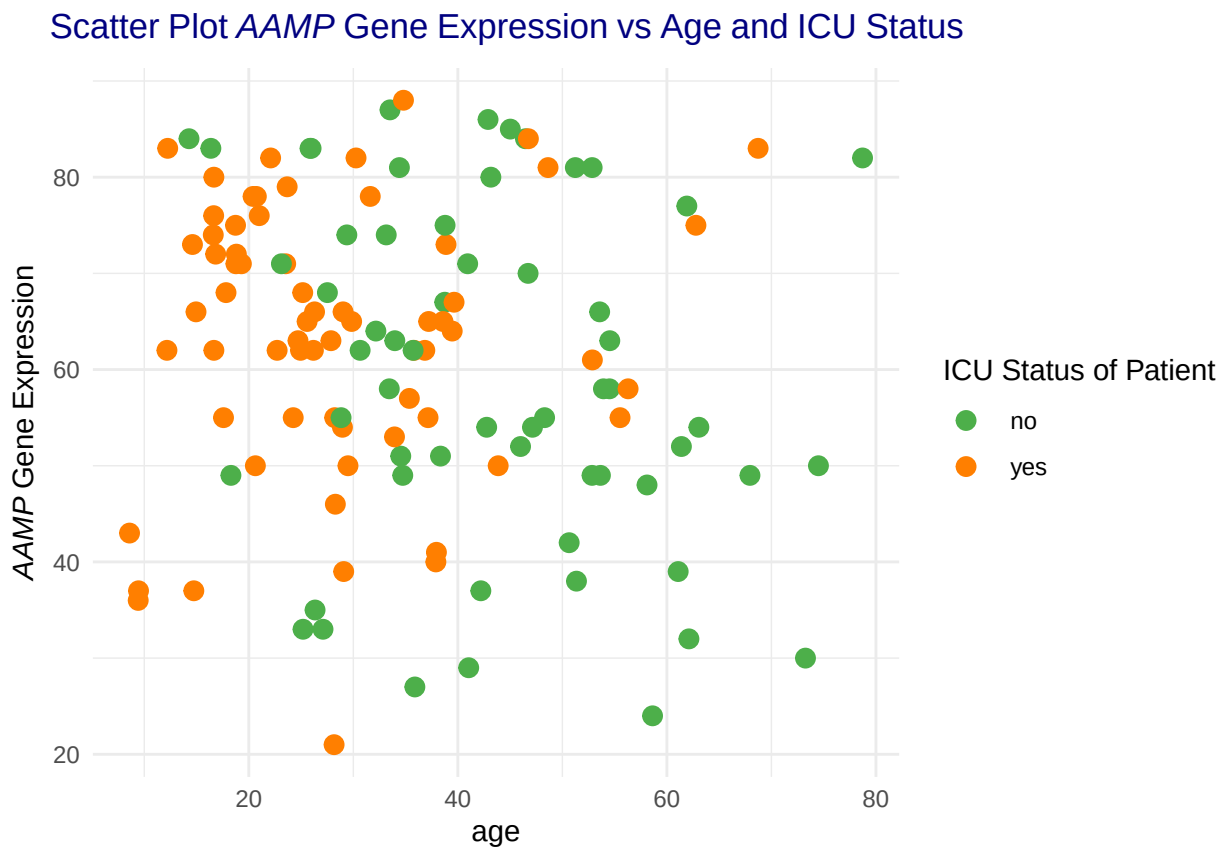
```
  print(all.plots.list[[gene]]$histogram)
  print(all.plots.list[[gene]]$scatter.plot)
  print(all.plots.list[[gene]]$box.plot)
}
```

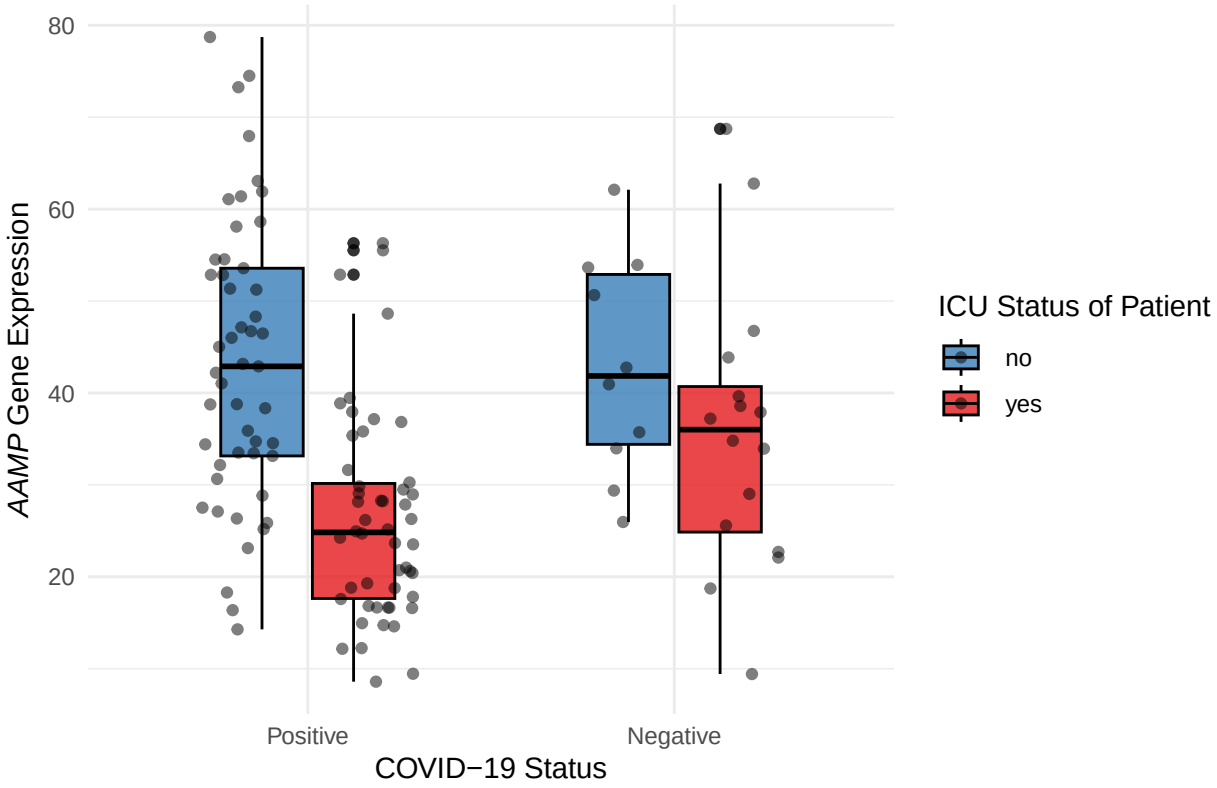


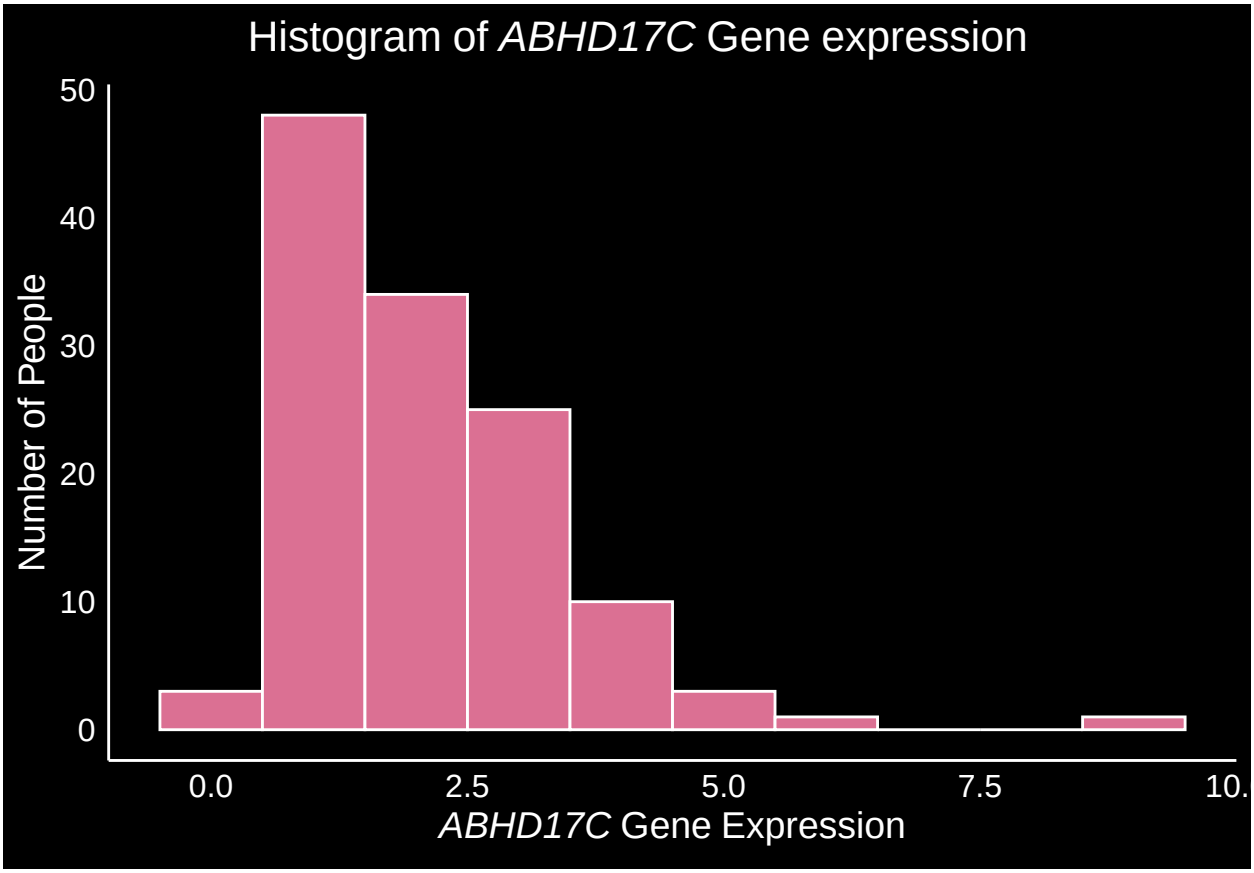

Scatter Plot ABHD18 Gene Expression vs Age and ICU Status
ABHD18 Gene Expression
80
60
40
20
2.5
5.0
7.5
10.0
age
ICU Status of Patient
no
yes

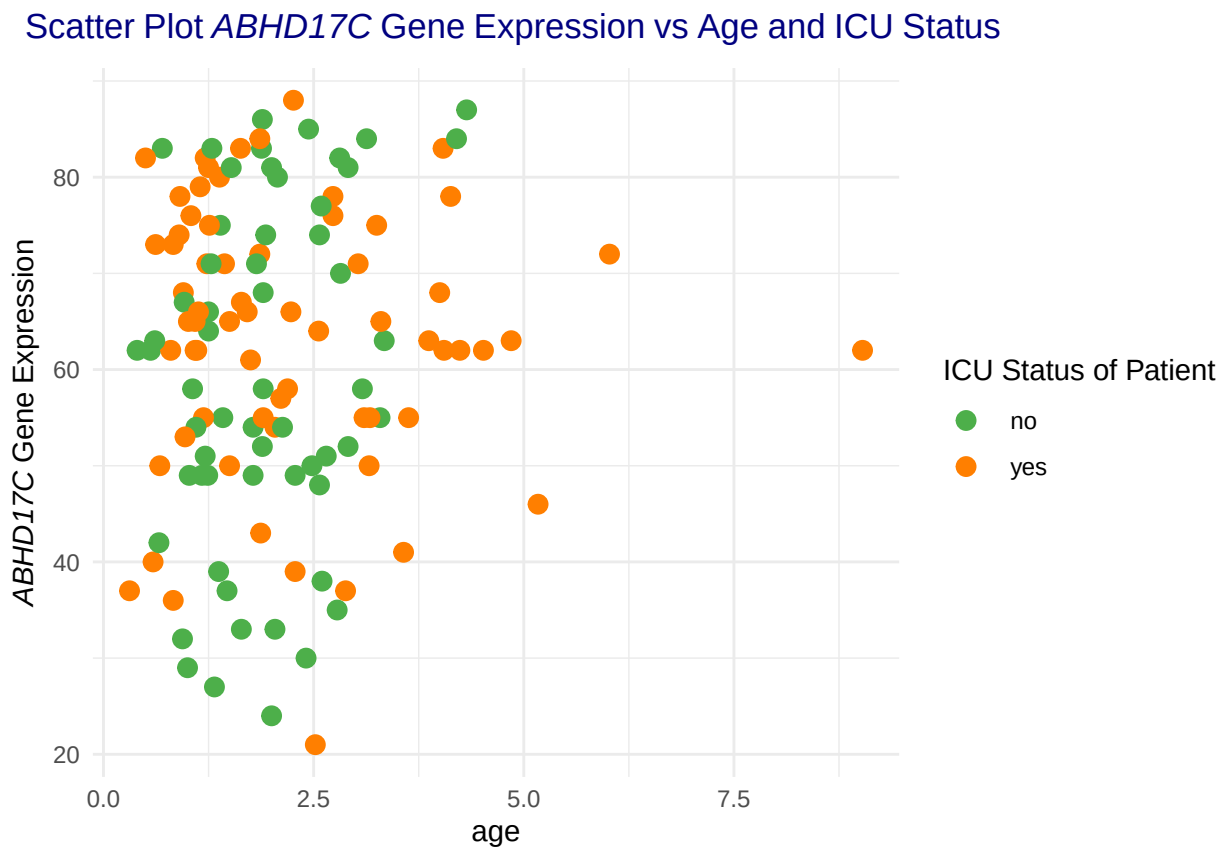

Box plot of ABHD18 Gene Expression by COVID and ICU Status
10.0
7.5
5.0
2.5
ABHD18 Gene Expression
Positive
Negative
COVID−19 Status
ICU Status of Patient
no
yes


Histogram of *AAMP* Gene expression
Number of People
0
2
4
6
20
40
60
80
*AAMP* Gene Expression


Scatter Plot AAMP Gene Expression vs Age and ICU Status
AAMP Gene Expression
80
60
40
20
20
40
60
80
age
ICU Status of Patient
no
yes


Box plot of AAMP Gene Expression by COVID and ICU Status
80
60
40
20
AAMP Gene Expression
Positive
Negative
COVID−19 Status
ICU Status of Patient
no
yes


Histogram of *ABHD17C* Gene expression
Number of People
50
40
30
20
10
0
0.0
2.5
5.0
7.5
10.
*ABHD17C* Gene Expression


Scatter Plot ABHD17C Gene Expression vs Age and ICU Status
ABHD17C Gene Expression
80
60
40
20
0.0
2.5
5.0
7.5
age
ICU Status of Patient
no
yes

Box plot of *ABHD17C* Gene Expression by COVID and ICU Status