Submisson2

Ethan Zhang

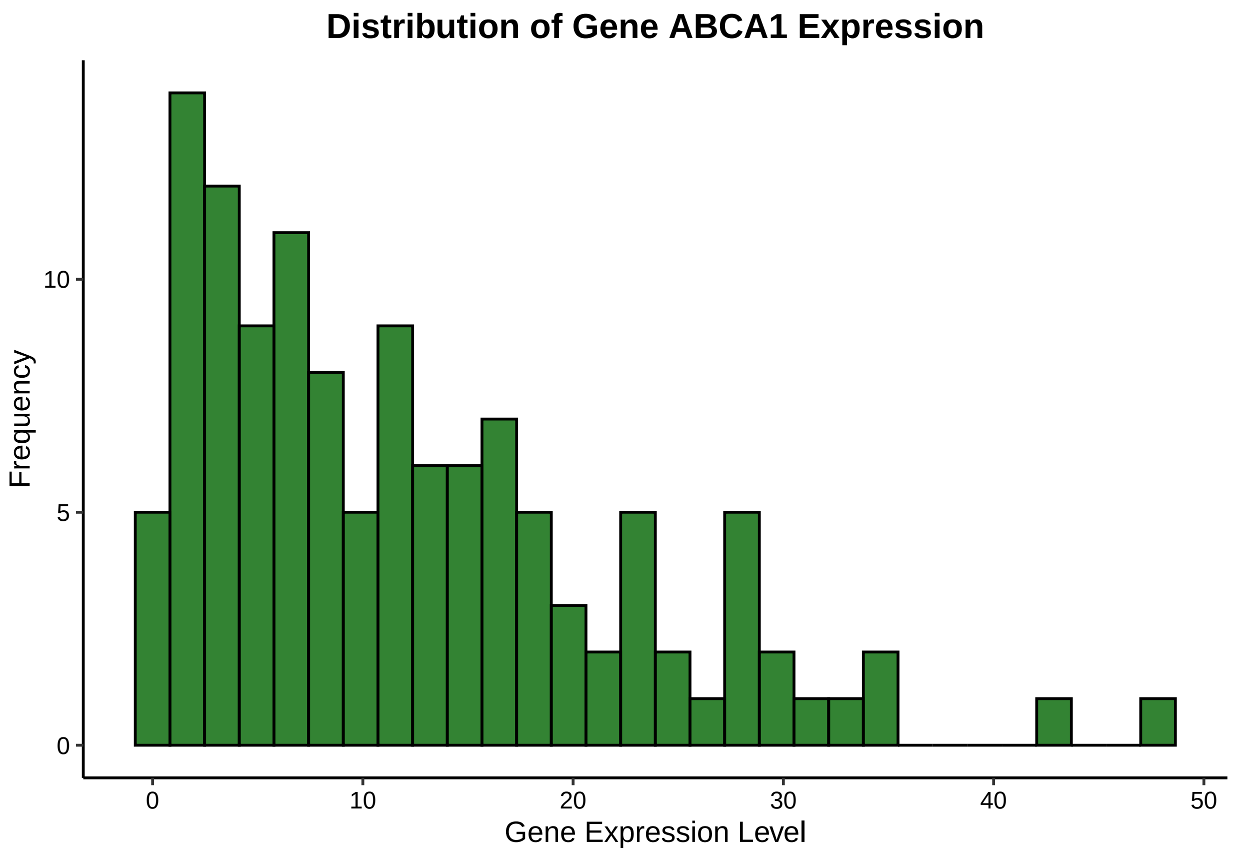
2024-08-07

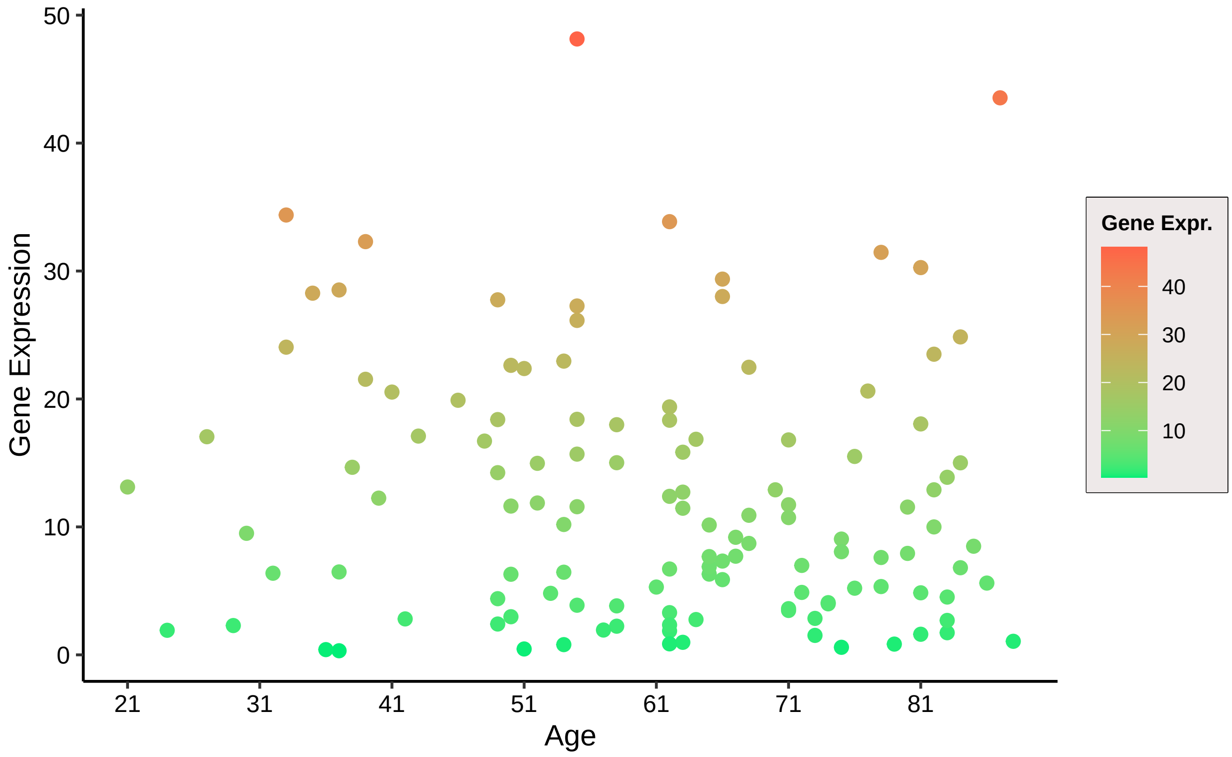
R Markdown

```r
#define a theme for all plots
New_theme <- theme(
  panel.border = element_blank(),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  # Set plot background
  plot.background = element_rect(fill = "white"),
  panel.background = element_blank(),
  legend.background = element_rect(fill = 'snow2'),
  legend.text = element_text(color = "black", size = 8),
  legend.title = element_text(color = "black", face = "bold", size = 8),
  legend.key = element_rect(fill = "snow2",color = "snow2"),
  legend.box.background  = element_rect(color = "black"),
  ##make the title center
  plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
  plot.subtitle = element_text(hjust = 0.5, size = 12, face = "italic"),
  title = element_text(color = "black"),
  axis.line = element_line(color = "black"),
  axis.text = element_text(color = "black"),
  legend.position = 'right'
)
```

```r
new_gene <-
  read.csv(
    "/Users/zhyihan/Documents/Dartmouth Life/courses/QBS 103-Data Science/final project/QBS103_GSE157103
  )
ser_m <-
  read.csv(
    "/Users/zhyihan/Documents/Dartmouth Life/courses/QBS 103-Data Science/final project/QBS103_GSE157103
  )

new_gene_long <-
  new_gene %>% gather(
    key = participant_id,
    value = expression,
    COVID_01_39y_male_NonICU:NONCOVID_26_36y_male_ICU
  )
```

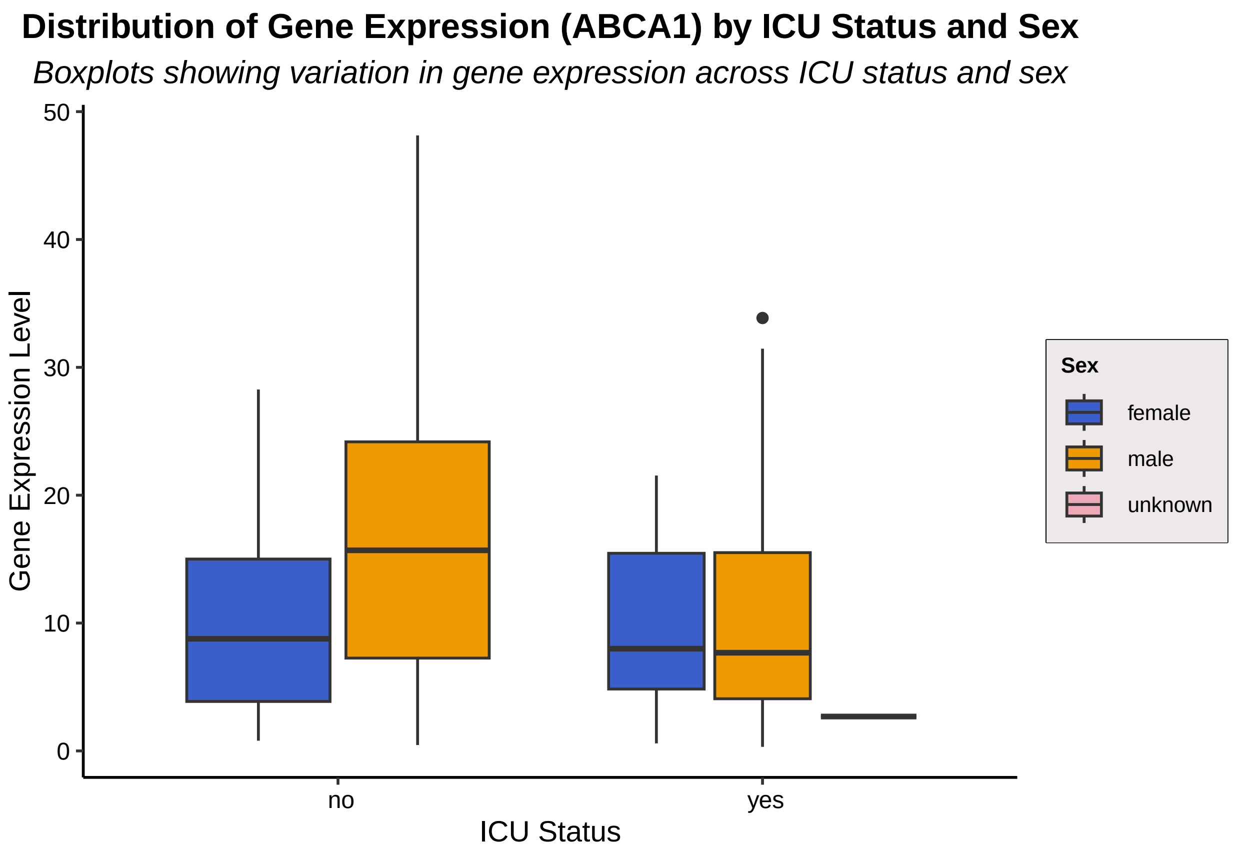
```r
new_data <- merge(ser_m, new_gene_long, by = "participant_id")


# Define the function
###replace the variable in our submission 1 as the variable names in our function now
###input data frame df; a list of gene: genes; continous variable: cont; two categorical varialbes: cate
gene_plots <- function(df, genes, cont, cate1, cate2) {
  df[[cont]] <- as.numeric(as.character(df[[cont]]))

  for (gene in genes) {
    new_data2 <- df %>%
      filter(X == gene) %>%
      select(participant_id,
             X,
             expression,
             cont,
             cate1,
             cate2) %>%
      drop_na()

    ####Histogram for gene expression
    hist <-
      ggplot(new_data2, aes_string(x = "expression")) + geom_histogram(
        bins = 30,
        fill = "darkgreen",
        color = "black",
        alpha = 0.8
      ) + labs(
        title = paste0("Distribution of Gene ", gene, " Expression"),
        x = "Gene Expression Level",
        y = "Frequency"
      ) + New_theme


    scatter_plot <- ggplot(new_data2,
                           aes_string(x = cont, y = "expression", color = "expression")) +
      geom_point(size = 2) +
      labs(
        title = paste0("Relationship Between ", gene, " Expression and ", cont),
        subtitle = paste0(
          "A scatterplot of gene expression levels across different ",
          cont
        ),
        y = 'Gene Expression',
        x = 'Age',
        color = 'Gene Expr.'
      ) +
      scale_color_gradient(low = "springgreen2", high = "tomato1") + New_theme +
      scale_x_continuous(breaks = seq(
        min(new_data2[[cont]], na.rm = TRUE),
        max(new_data2[[cont]], na.rm = TRUE),
        by = 10
      ))
    ####Boxplot of gene expression separated by both categorical covariates
    new_scat <-
      ggplot(new_data2,
             aes_string(x = cate2, y = "expression", fill = cate1)) +
      geom_boxplot() +
      New_theme +
      labs(
        title = paste0(
          "Distribution of Gene Expression (",
          gene,
          ") by ICU Status and Sex"
        ),
        subtitle = "Boxplots showing variation in gene expression across ICU status and sex",
        x = "ICU Status",
        y = "Gene Expression Level",
        fill = "Sex"
      )+scale_fill_manual(values = c('royalblue3', 'orange2','pink2'))

    print(hist)
    print(scatter_plot)
    print(new_scat)
  }
}

gene_plots(new_data, list("ABCA1", "AATF","A2M"), "age", "sex", "icu_status") %>% suppressWarnings()
```

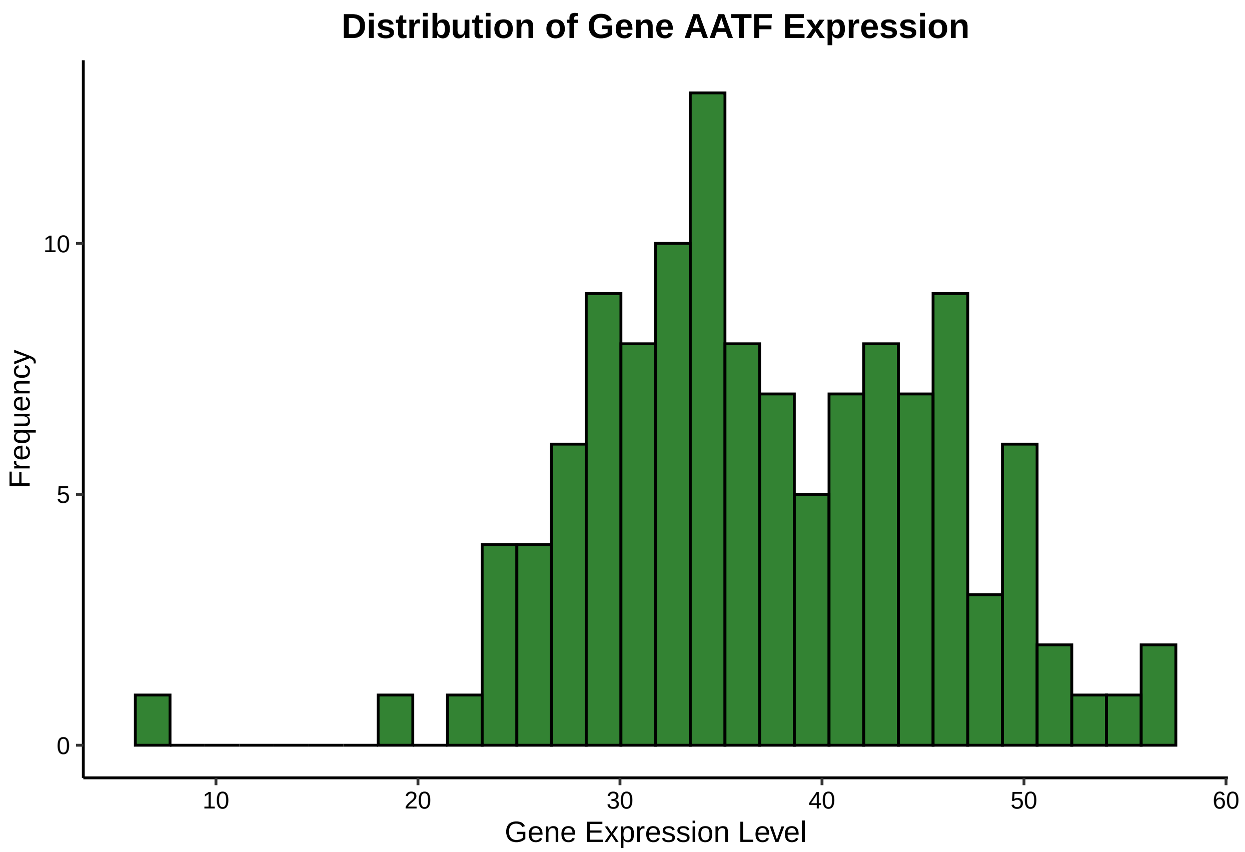

Distribution of Gene ABCA1 Expression
Frequency
10
5
0
0
10
20
30
40
50
Gene Expression Level

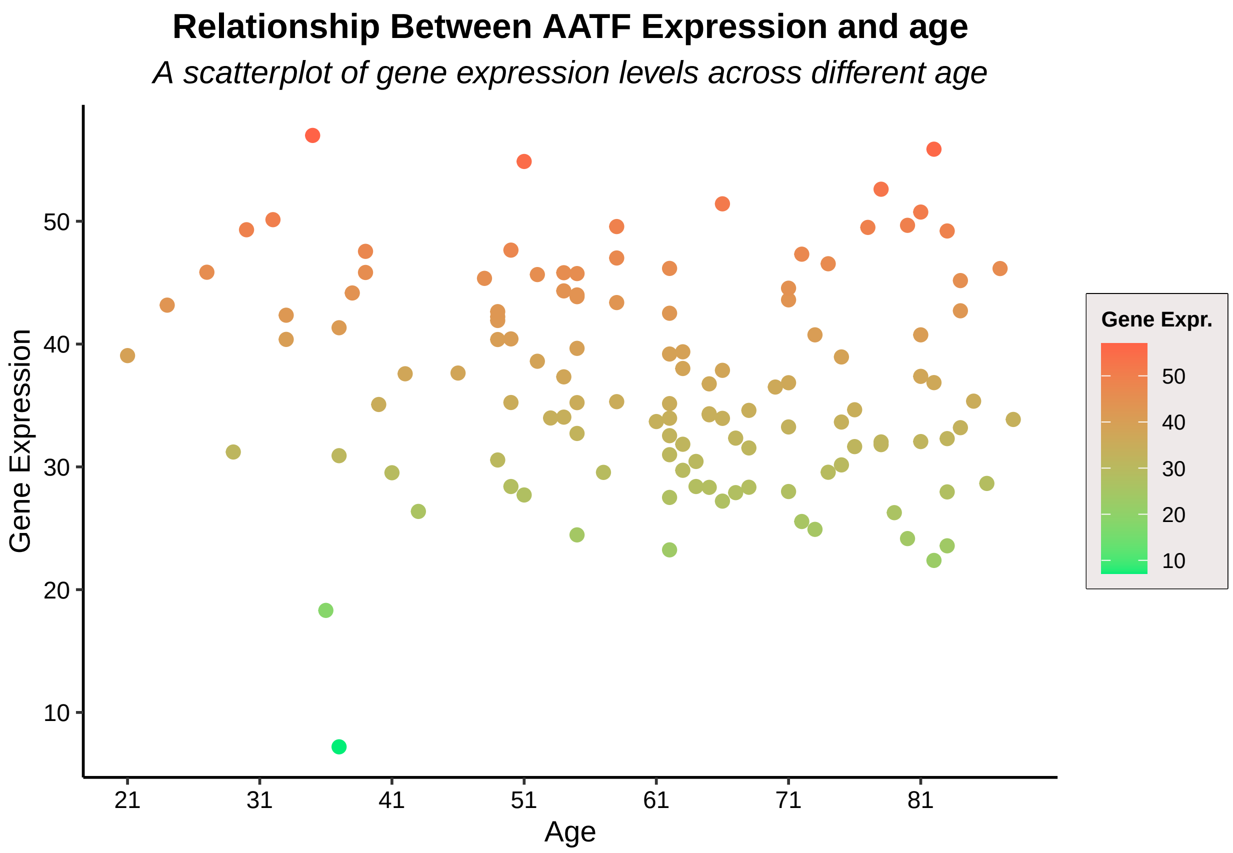


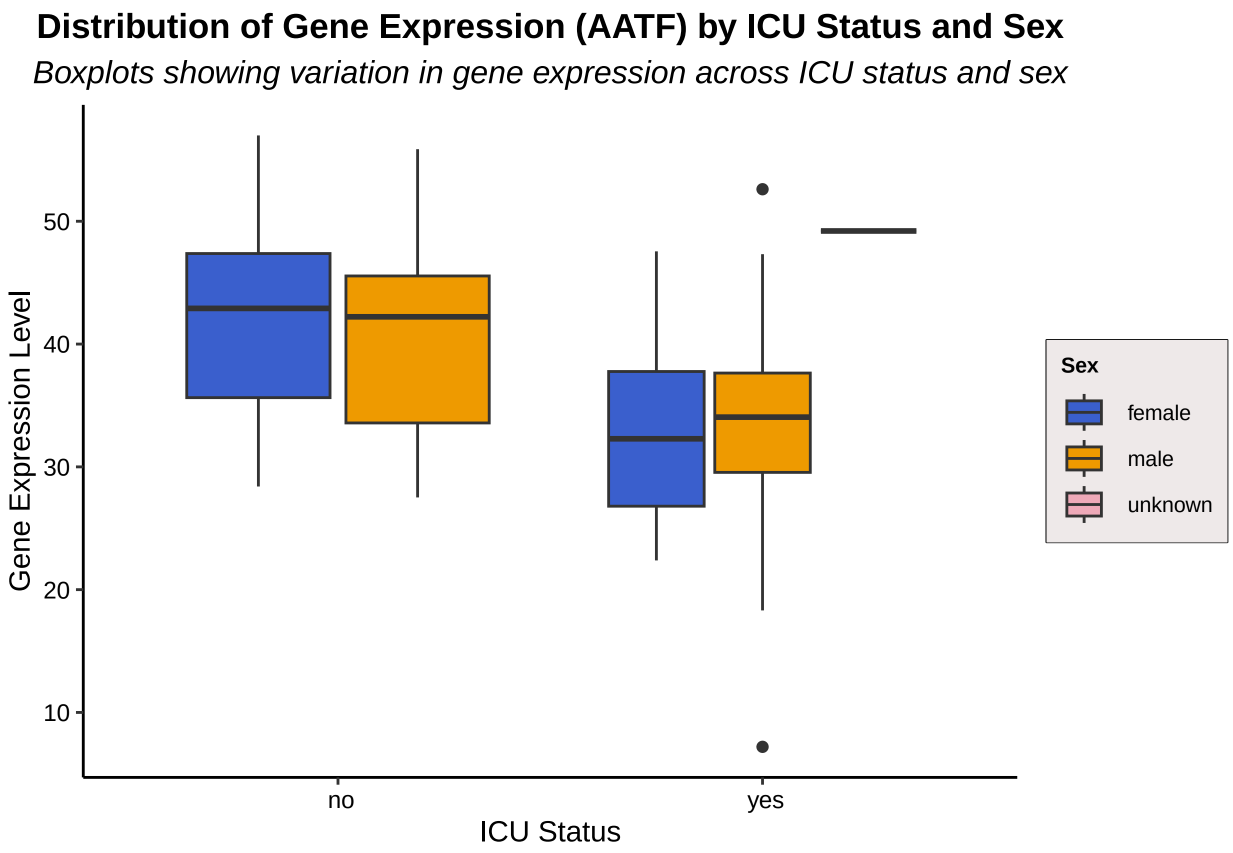
Relationship Between ABCA1 Expression and age

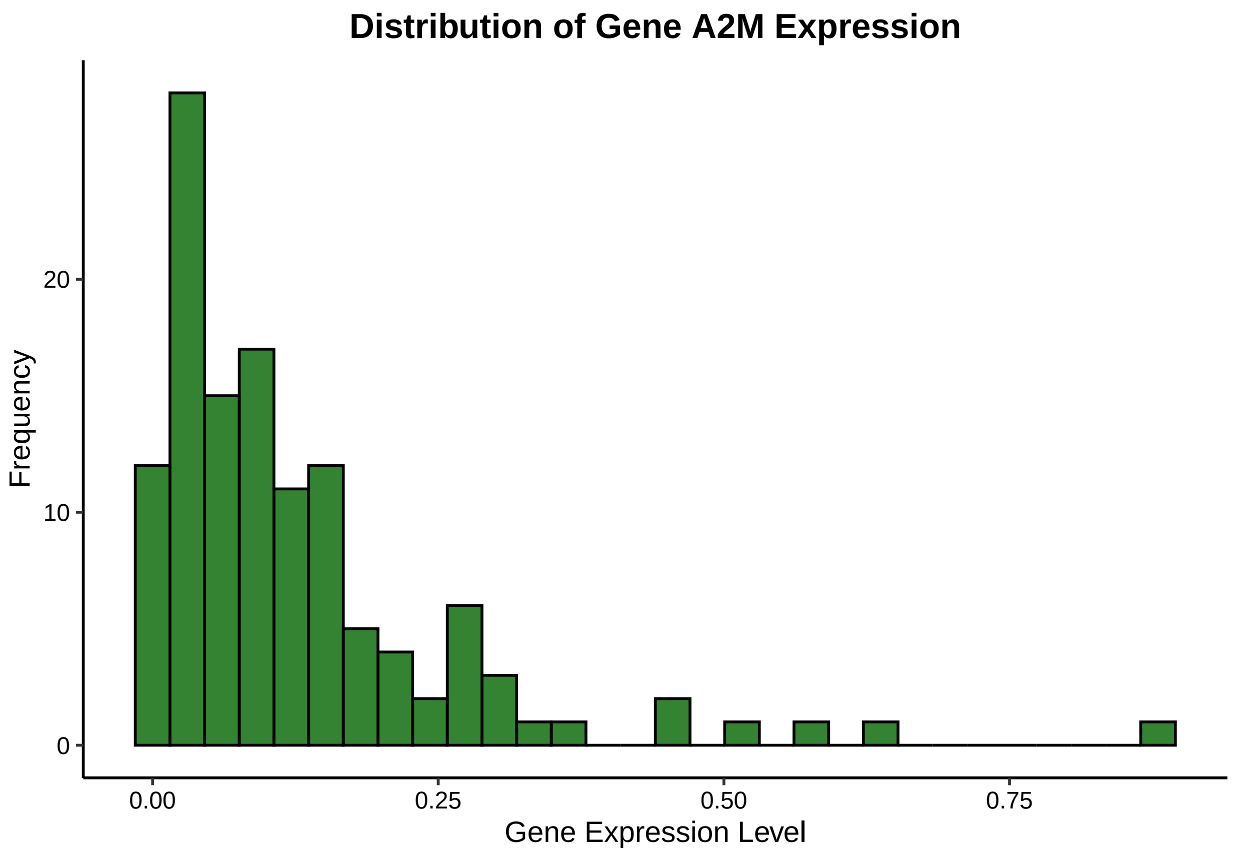
A scatterplot of gene expression levels across different age

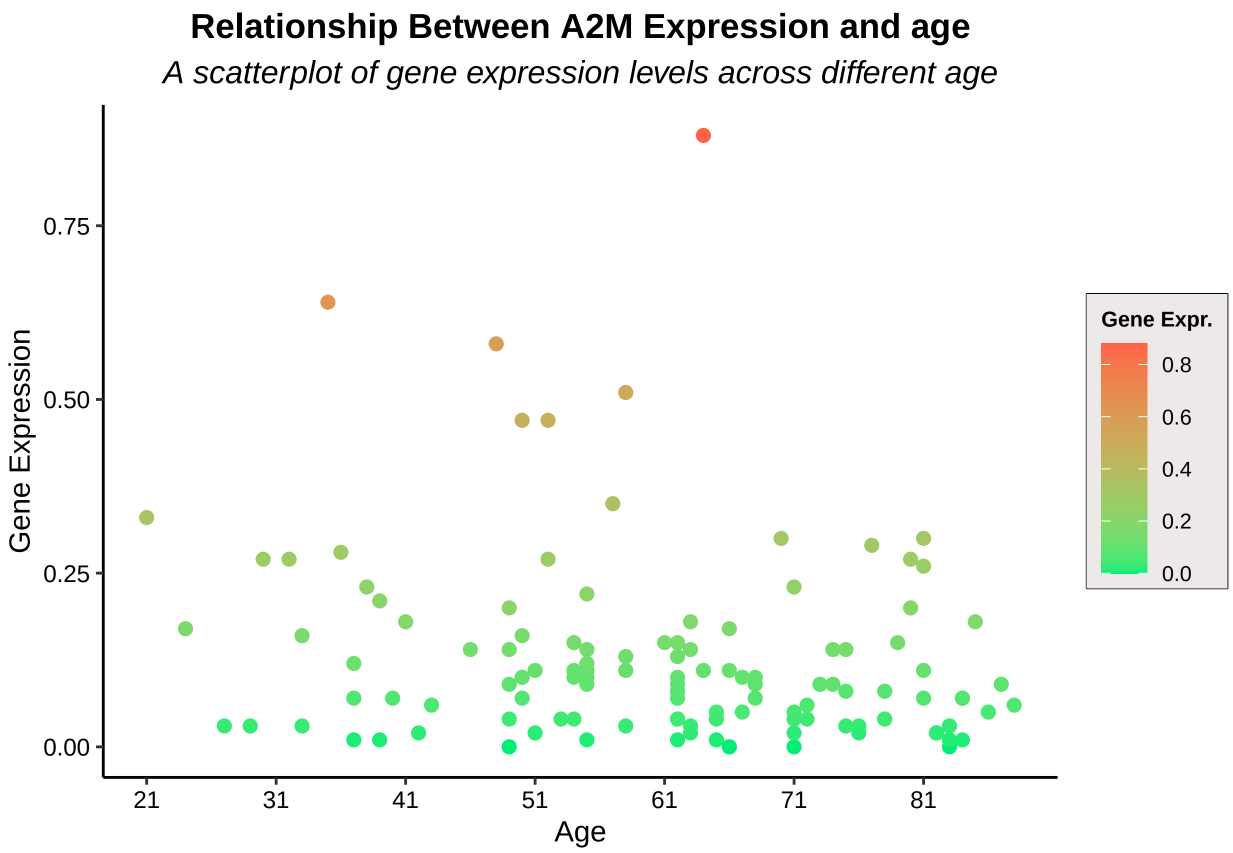

Distribution of Gene Expression (ABCA1) by ICU Status and Sex
Boxplots showing variation in gene expression across ICU status and sex
Gene Expression Level
50
40
30
20
10
0
no
yes
ICU Status
Sex
female
male
unknown

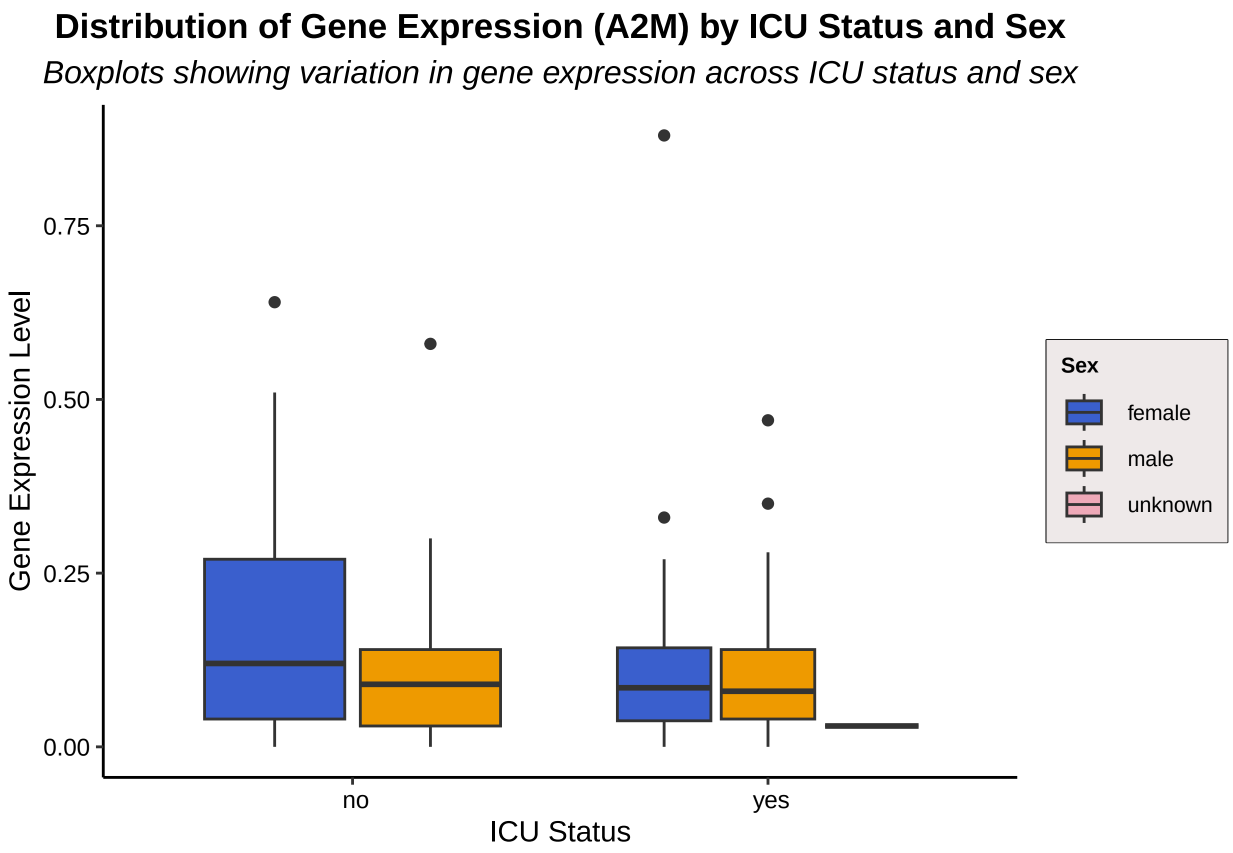

Distribution of Gene AATF Expression
Frequency
0
5
10
10
20
30
40
50
60
Gene Expression Level


Relationship Between AATF Expression and age
A scatterplot of gene expression levels across different age
Gene Expression
50
40
30
20
10
21
31
41
51
61
71
81
Age
Gene Expr.
50
40
30
20
10


Distribution of Gene Expression (AATF) by ICU Status and Sex
Boxplots showing variation in gene expression across ICU status and sex
Gene Expression Level
50
40
30
20
10
no
yes
ICU Status
Sex
female
male
unknown


Distribution of Gene A2M Expression
Frequency
20
10
0
0.00
0.25
0.50
0.75
Gene Expression Level


Relationship Between A2M Expression and age
A scatterplot of gene expression levels across different age
Gene Expression
0.75
0.50
0.25
0.00
21
31
41
51
61
71
81
Age
Gene Expr.
0.8
0.6
0.4
0.2
0.0



Distribution of Gene Expression (A2M) by ICU Status and Sex

Boxplots showing variation in gene expression across ICU status and sex