Project 1_submission2

Luwei Zeng

2024-08-06

Build a function to create the plots you made for Presentation 1,incorporating any feedback you received on your submission. Your functions should take the following input: (1) the name of the data frame, (2) a list of 1 or more gene names, (3) 1 continuous covariate, and (4) two categorical covariates

Select 2 additional genes (for a total of 3 genes) to look at and implement a loop to generate your figures using the function you created (10 pts)

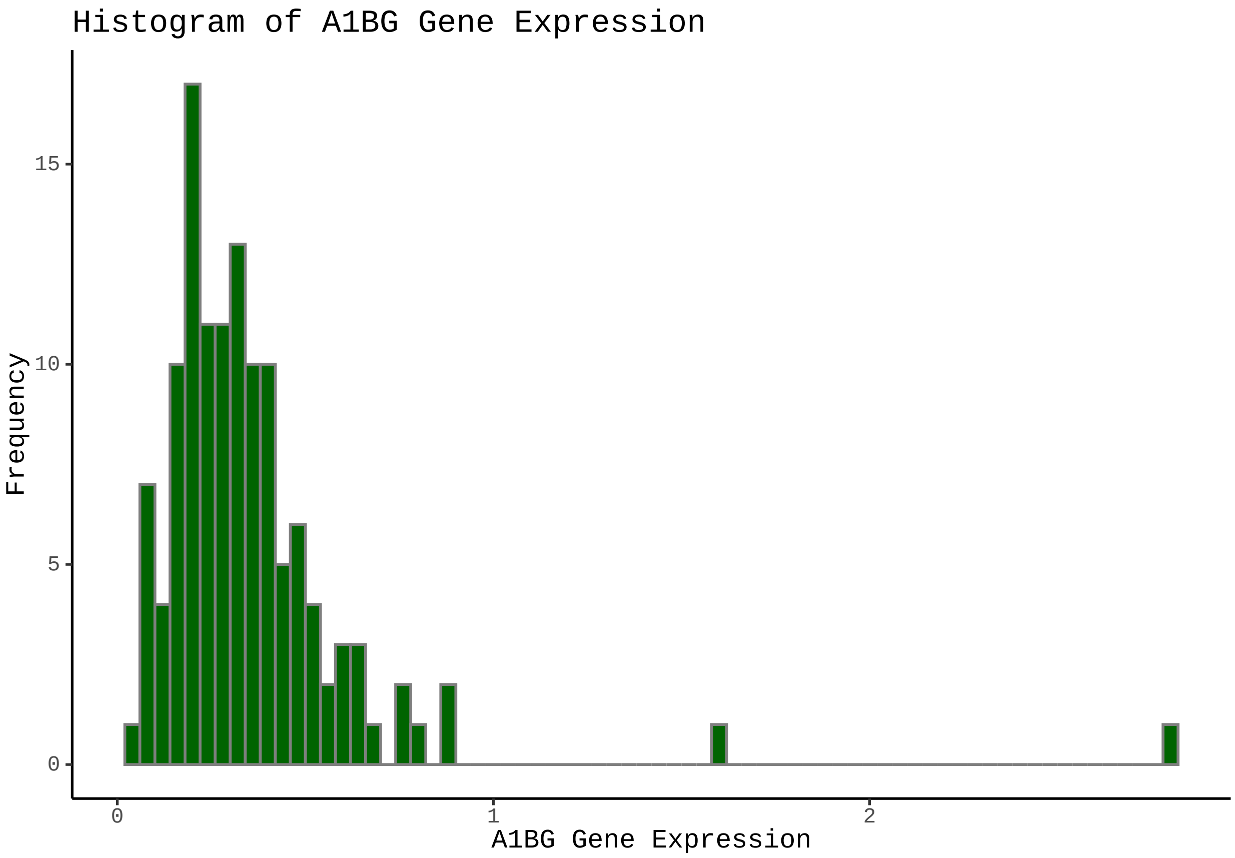
```
library(tidyverse)
```

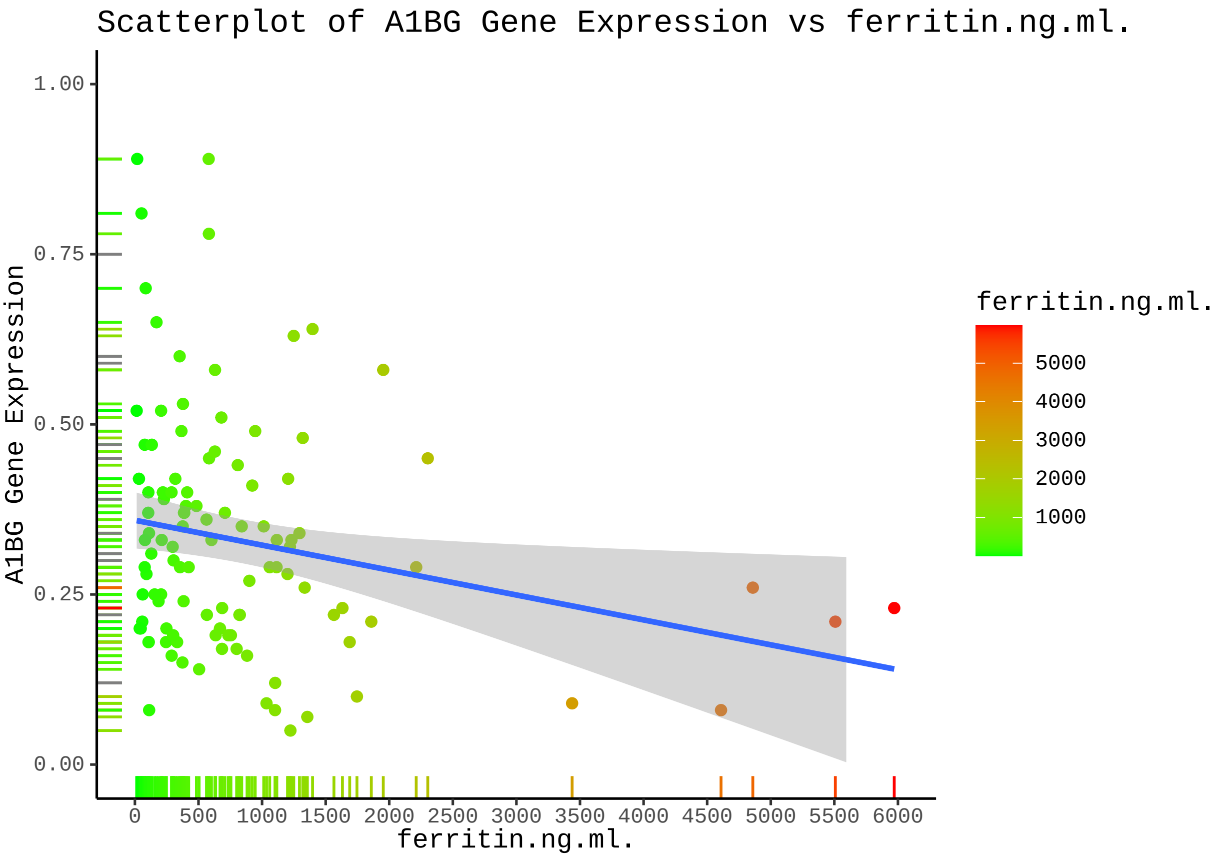
One gene:A1BG; Three genes:A1BG,A4GNT,and A1CF; One continuous covariate:ferritin(ng/ml); Two categorical covariates:sex, icu status.

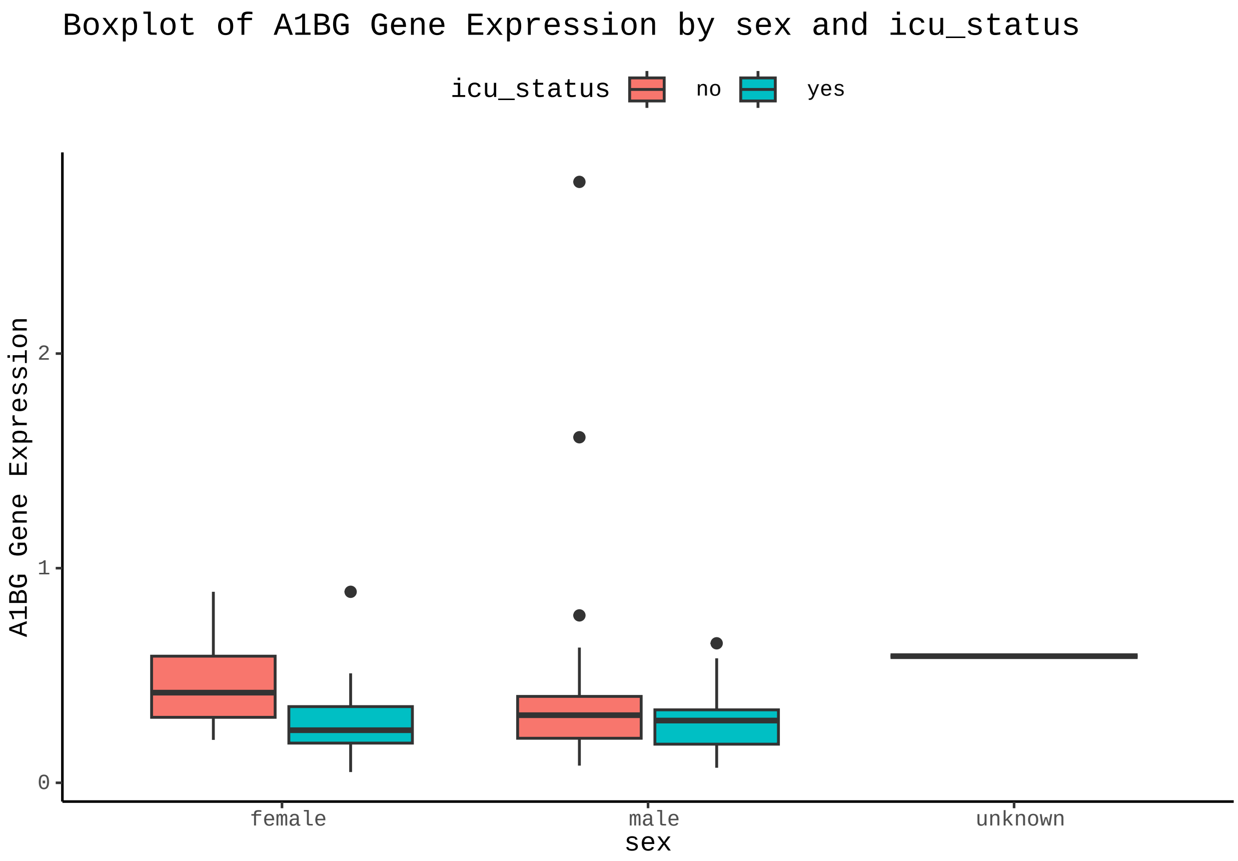
```
#Load data
setwd("/Users/hekaiwei/Desktop/R class/project")
gene <- read.csv(file = "QBS103_GSE157103_genes.csv",row.names=1)
metadata <- read.csv(file = "QBS103_GSE157103_series_matrix.csv", row.names = 1)

#convert to dataframe and switch column and row for multiple genes
selected_genes<-c("A1BG","A4GNT","A2M") #select "A1BG,A4GNT,A2M"
gene_data<-gene[selected_genes,]
gene_data <- as.data.frame(t(gene_data)) #convert to dataframe and switch column and row

#combine selected genes expression and metadata
data_combined<-merge(gene_data,metadata,by = "row.names")
colnames(data_combined)[1] <- "ID" #set the 1st column name as "ID"
```

```
#Create a function from the datasource that include gene expression, one continuous covariate,
#and two categorical covariates
plot_created<-function(data_source,gene_names,
                       continuous_covariate,
                       categorical_covariate1,categorical_covariate2){

  for (gene_name in gene_names){
      #Ensure that continuous covariate is numeric
      data_source[[continuous_covariate]] <- as.numeric(data_source[[continuous_covariate]])


      #Histogram for for gene expression
      plot1 <- ggplot(data_source, aes_string(x = gene_name)) +
      #Add histogram data with specific bin width, fill color, and border color
      geom_histogram(binwidth=0.04, fill="darkgreen",color="grey50")+
      labs(title = paste("Histogram of", gene_name, "Gene Expression"),
        x=paste(gene_name,"Gene Expression"),
        y="Frequency")+
      theme_classic(base_family = 'Courier',base_size = 10)

      print(plot1)

      #Scatterplot for gene expression and one continuous covariate
      plot2<-ggplot(data_source, aes_string(x =continuous_covariate,
                                            y = gene_name,
                                            color=continuous_covariate)) +
      #Add points to the plot, setting the color of the points to dark green
      geom_point() +
      geom_smooth(method=lm)+ #add trendline
      geom_rug(sides="bl")+ #visualize the density of the data
      labs(title=paste("Scatterplot of",gene_name,"Gene Expression vs", continuous_covariate),
        x = continuous_covariate,
        y = paste(gene_name,"Gene Expression"))+
      ylim(0,1)+
      scale_x_continuous(limits = c(0, 6000), breaks = seq(0, 6000, by = 500)) +
      #add color gradient
      scale_color_gradient(low = "green", high = "red") +
      theme_classic(base_family = 'Courier',base_size = 10)

      print(plot2)

      #Boxplot for gene expression by categorical covariates
      plot3<-ggplot(data_source, aes_string(x = categorical_covariate1,
                                            y = gene_name, fill = categorical_covariate2)) +
      #Add boxplot
      geom_boxplot()+
      labs(title = paste("Boxplot of",gene_name,"Gene Expression by",
                         categorical_covariate1, "and",categorical_covariate2),
        x = categorical_covariate1,
        y = paste(gene_name,"Gene Expression"),
        fill = categorical_covariate2)+
      theme_classic(base_family = 'Courier',base_size = 10)+
      #Customize the theme to position the legend at the top of the plot
      theme(legend.position = 'top')

      print(plot3)
  }
}
```

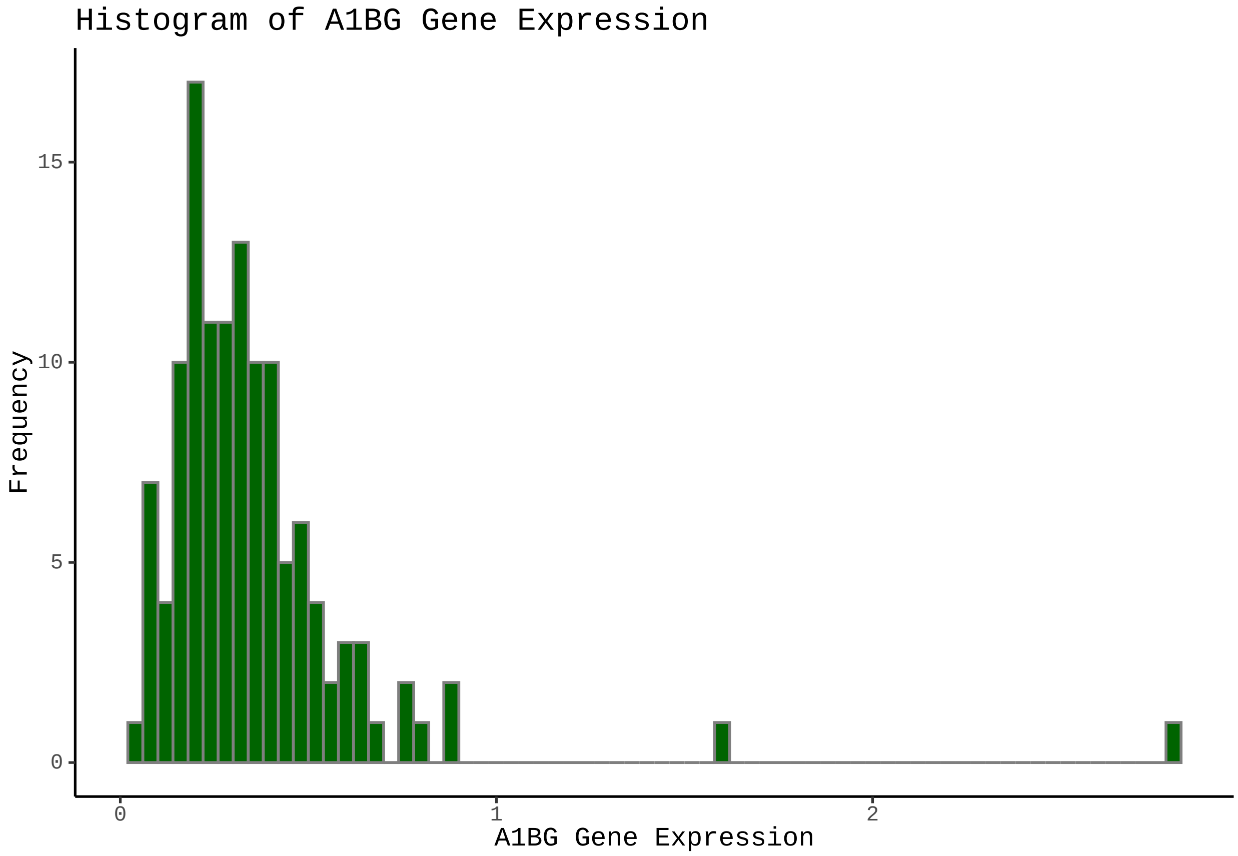
```
#Generate plots for A1BG by calling the function
plot_created(data_source = data_combined,
             gene_names = 'A1BG',
             continuous_covariate = "ferritin.ng.ml.",
             categorical_covariate1 = 'sex',
             categorical_covariate2 = 'icu_status')
```

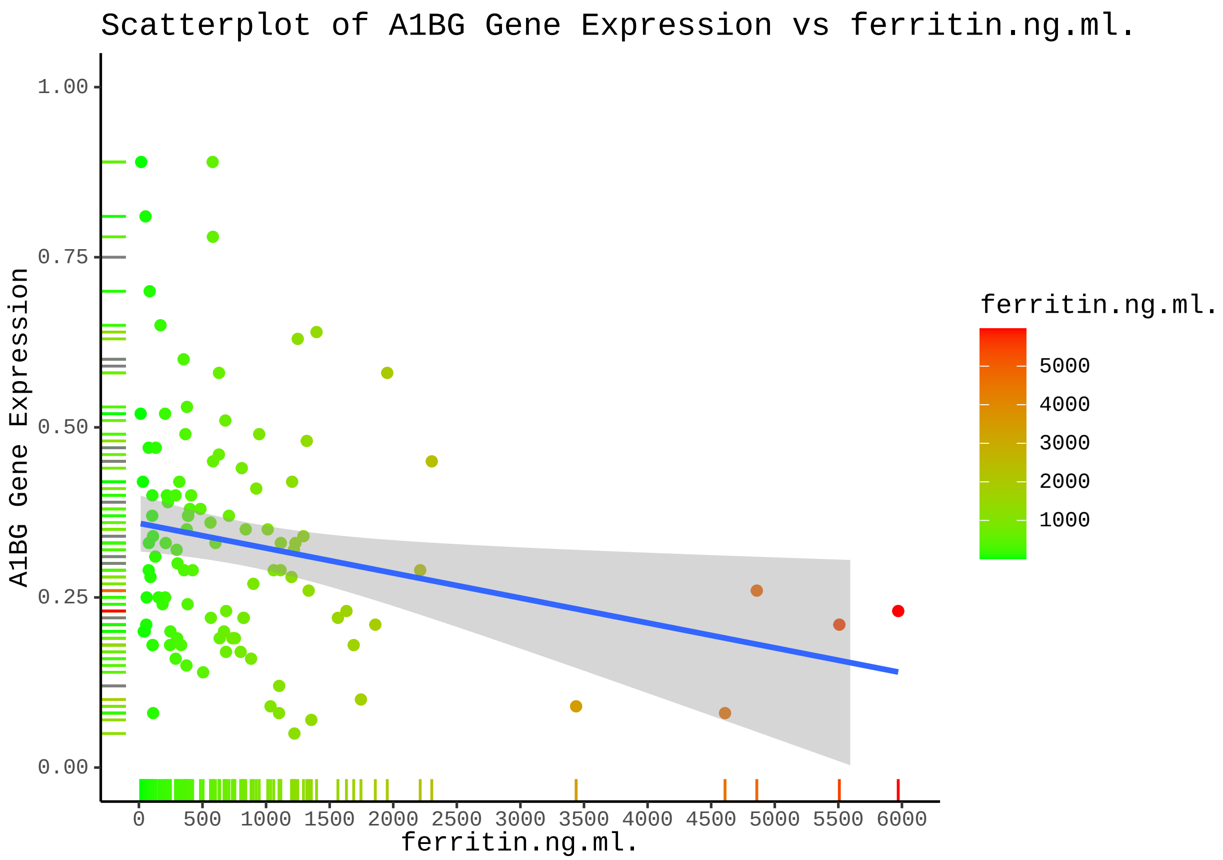

Histogram of A1BG Gene Expression
Frequency
0
5
10
15
0
1
2
A1BG Gene Expression

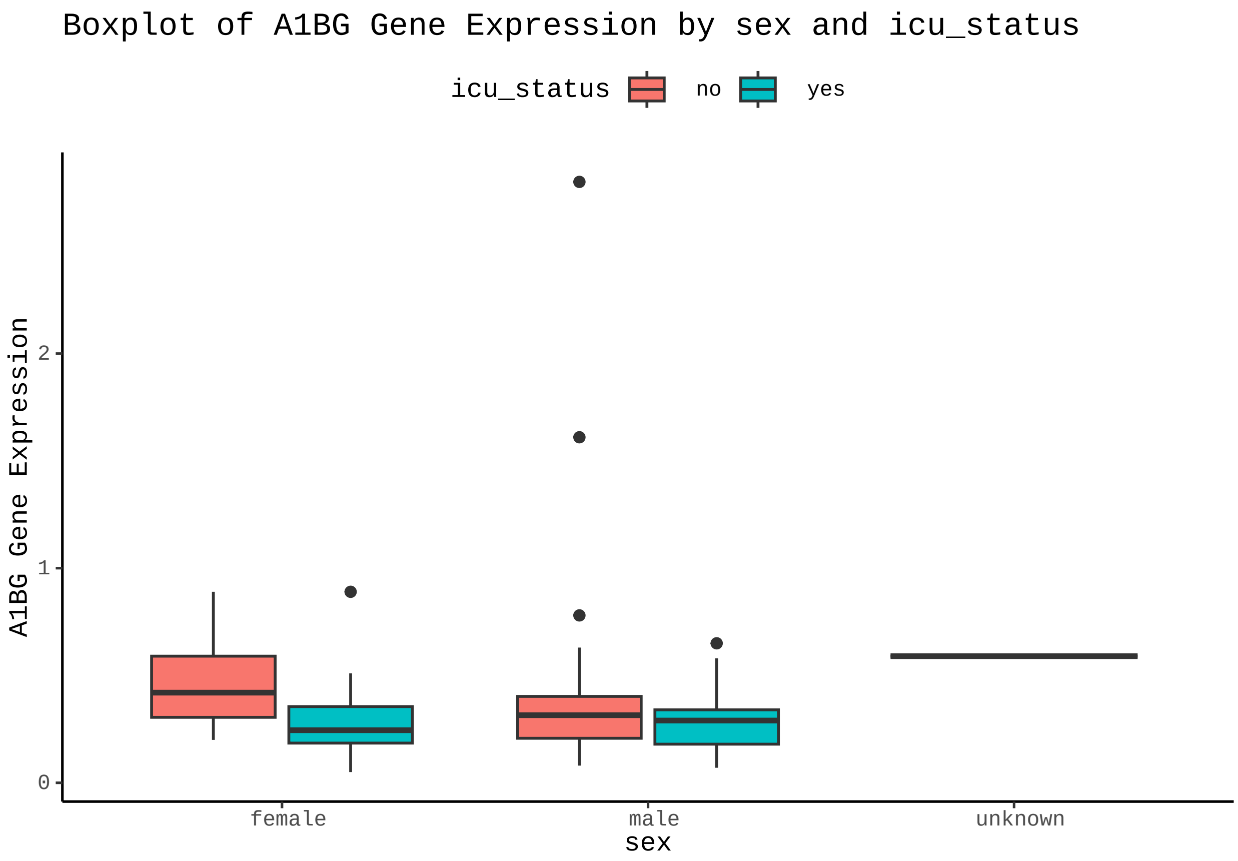

Scatterplot of A1BG Gene Expression vs ferritin.ng.ml.
A1BG Gene Expression
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
ferritin.ng.ml.
ferritin.ng.ml.
5000
4000
3000
2000
1000

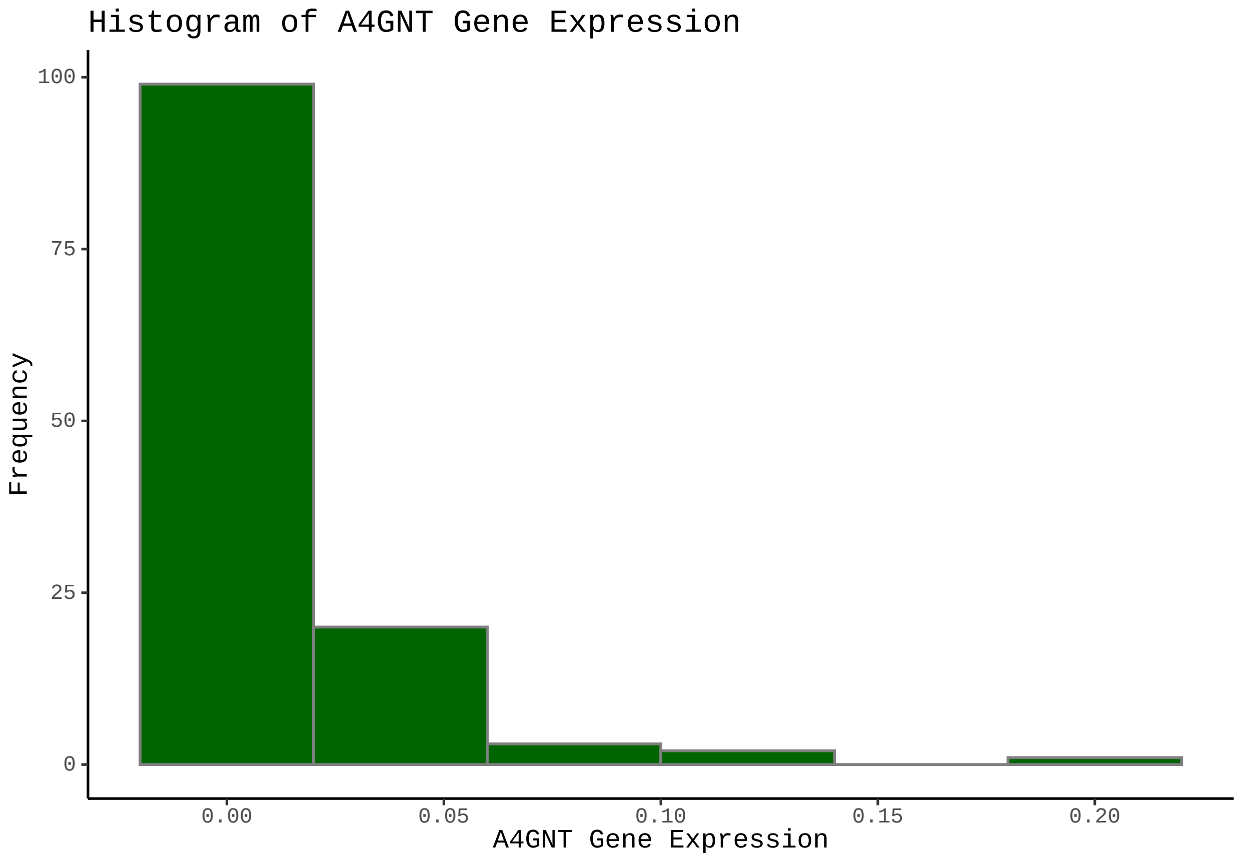


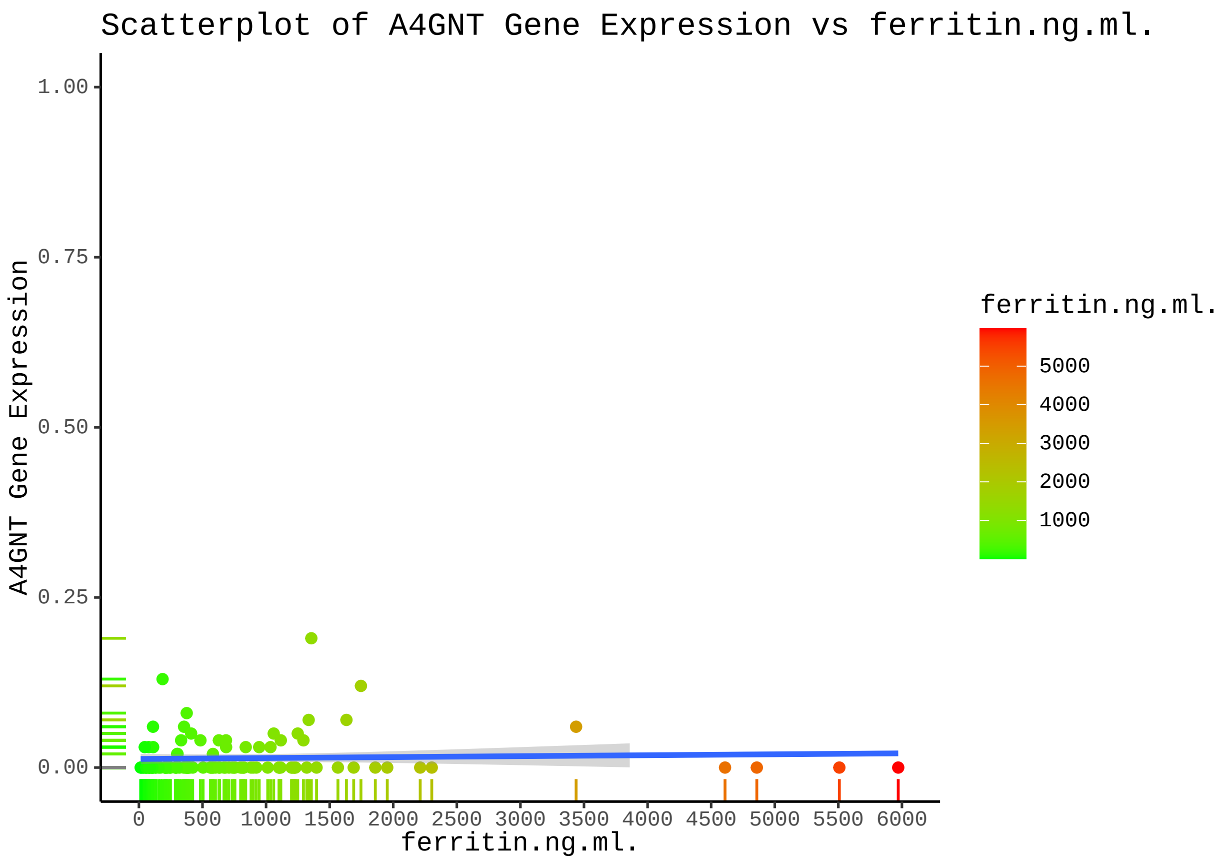
```r
#Generate plots for A1BG,A4GNT,and A1CF by calling the function
plot_created(data_source = data_combined,
             gene_names = selected_genes,
             continuous_covariate = "ferritin.ng.ml.",
             categorical_covariate1 = 'sex',
             categorical_covariate2 = 'icu_status')
```

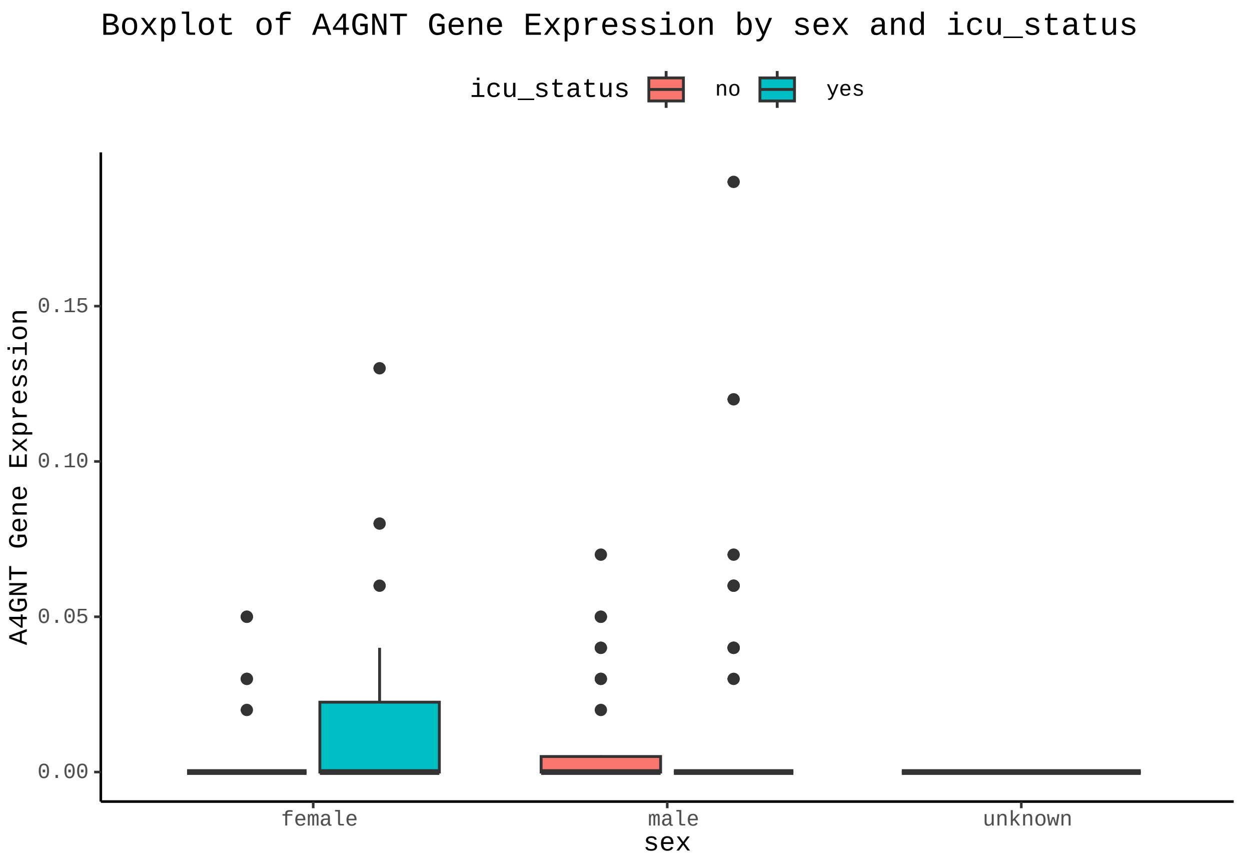

Histogram of A1BG Gene Expression
Frequency
15
10
5
0
0
1
2
A1BG Gene Expression

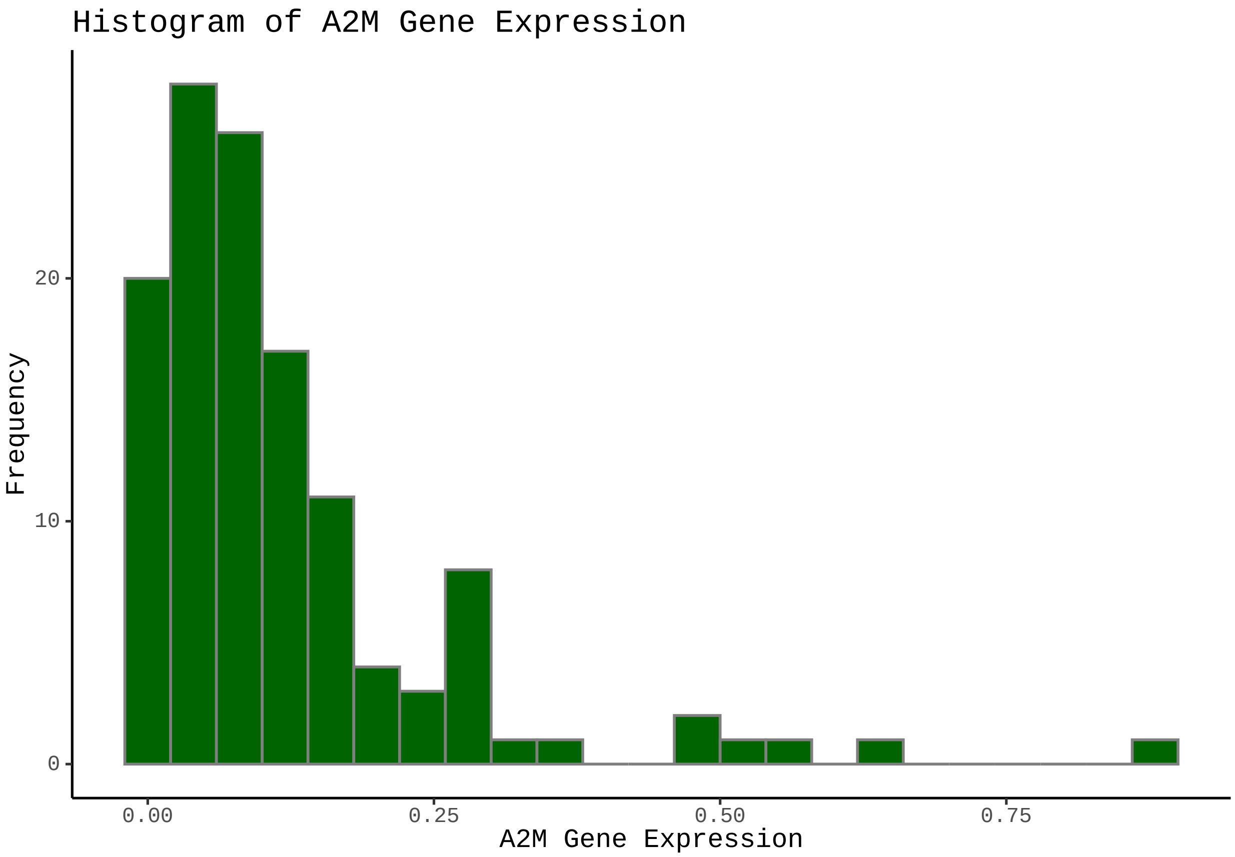

Scatterplot of A1BG Gene Expression vs ferritin.ng.ml.
A1BG Gene Expression
1.00
0.75
0.50
0.25
0.00
ferritin.ng.ml.
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
ferritin.ng.ml.
5000
4000
3000
2000
1000

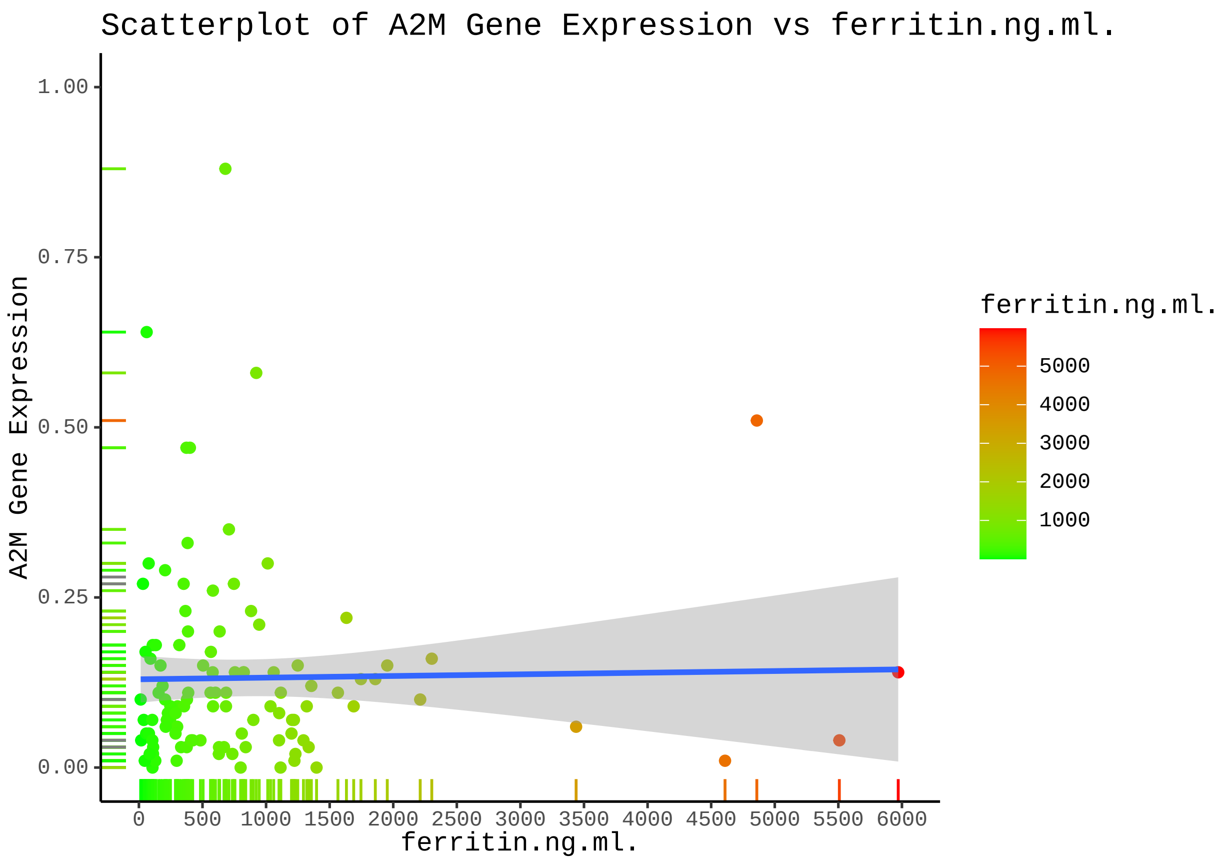

Boxplot of A1BG Gene Expression by sex and icu_status
icu_status
no
yes
A1BG Gene Expression
0
1
2
female
male
unknown
sex

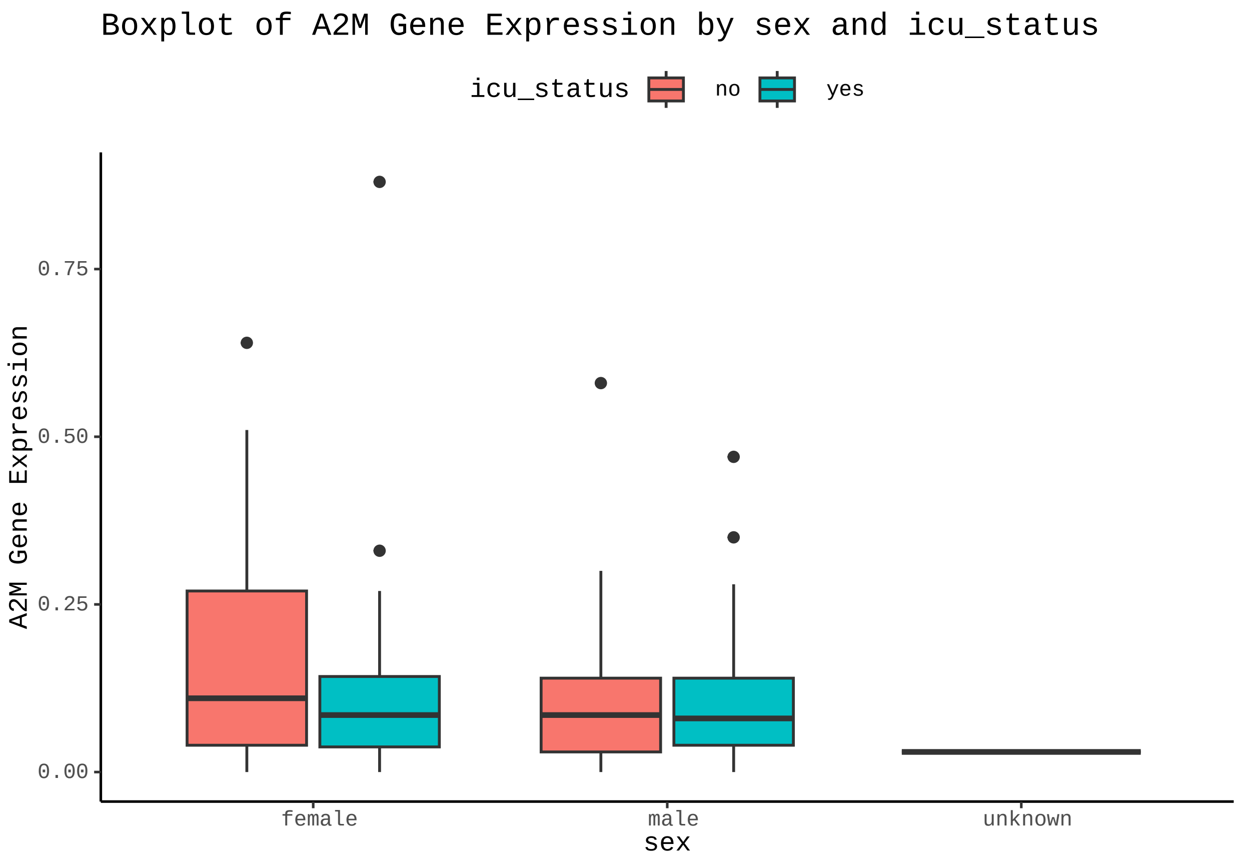

Histogram of A4GNT Gene Expression
100
75
50
25
0
Frequency
0.00
0.05
0.10
0.15
0.20
A4GNT Gene Expression


Scatterplot of A4GNT Gene Expression vs ferritin.ng.ml.
A4GNT Gene Expression
1.00
0.75
0.50
0.25
0.00
ferritin.ng.ml.
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
ferritin.ng.ml.
5000
4000
3000
2000
1000


Boxplot of A4GNT Gene Expression by sex and icu_status
icu_status
no
yes
A4GNT Gene Expression
0.15
0.10
0.05
0.00
female
male
unknown
sex


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


Scatterplot of A2M Gene Expression vs ferritin.ng.ml.
A2M Gene Expression
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
ferritin.ng.ml.
ferritin.ng.ml.
5000
4000
3000
2000
1000


Boxplot of A2M Gene Expression by sex and icu_status
icu_status
no
yes
A2M Gene Expression
0.75
0.50
0.25
0.00
female
male
unknown
sex