Submission_2_103

R Markdown

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 (10 pts) 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) Present one of your boxplots in class. Be prepared to explain the gene and covariates you chose and comment on the distribution as if you were presenting your research findings. No slides are required, just bring your plot. In class, be prepared to provide constructive feedback for your classmates (5 pts) Make sure you push your code to your git repository prior to class. As a reminder, we do not need you to share your GitHub repository until the final submission. Pushing this submission to GitHub will be worth 5 pts on the final submission and you can earn 1 additional point on your final project grade if you push 1 extra time along the way (changes between pushes must be significant to earn the extra point).

```
#Convert genes into gene series
library(readr)
setwd("/Users/deepthi/Documents/GitHub/Final_103_Project_LDG")
# read in genes file
Genes<- read_csv("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`
```

```
#read in Gene Series file
Genes_Series <- read_csv("QBS103_GSE157103_series_matrix.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
#transpose gene data to put participant id as rows and genes as columns
TGenes <-as.data.frame(t(Genes))
# rearrange table to make gene names as column names instead of row names
names(TGenes) <- TGenes[1,]
#remove repetitive first column with genes
TGenes <- TGenes[-1,]
#create a column with the row names( participant ids)to merge with gene series file
TGenes$participant_id <- row.names(TGenes)
# Merge Gene and Gene Series tables by participant_id
MergedGenes <- merge(TGenes,Genes_Series, by = 'participant_id', all = TRUE)
```

#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 (10 pts)

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

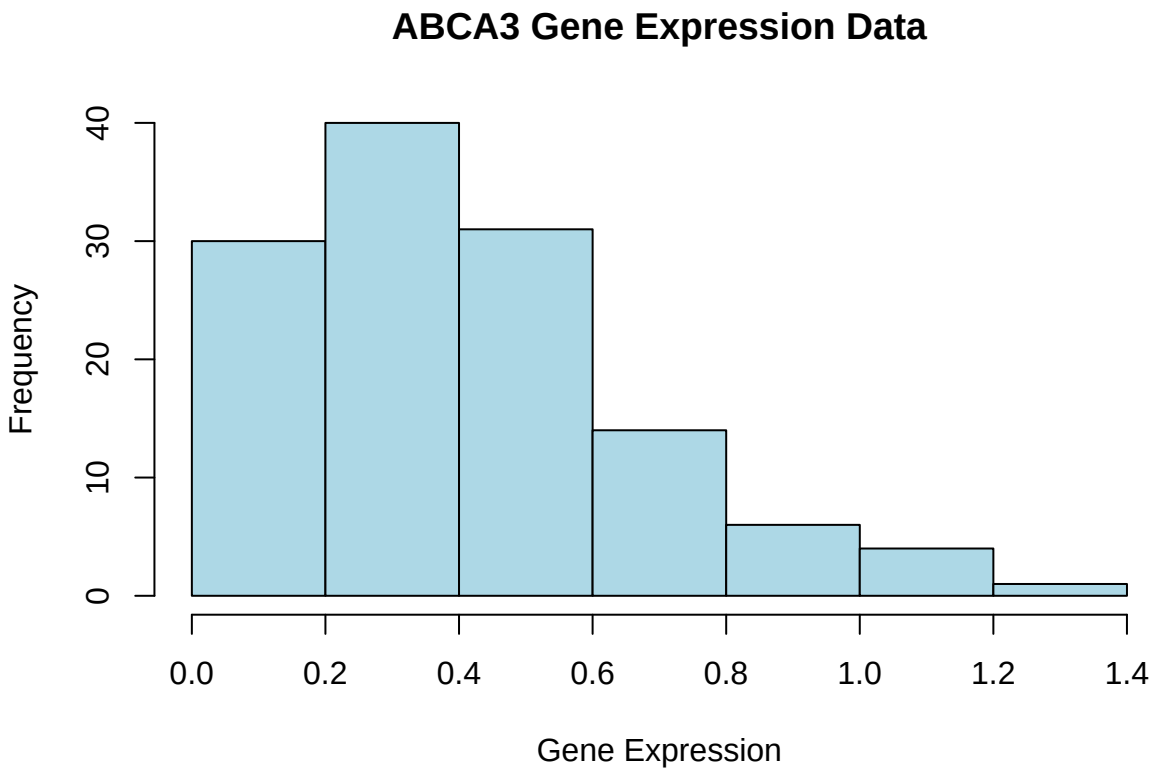
```r
library(ggplot2)

GenePlots <- function(df, geneName, Cont, Cat1, Cat2) {
  # Pull out gene expression data, continuous covariate, and categorical data
  geneName <- geneName[[1]]  # Assumes geneName is a single-element list
  Genes_CoVariate_Data <- df %>%
    select(all_of(c(geneName, Cont, Cat1, Cat2))) %>%
    as.data.frame()
  # Convert gene expression column to numeric
  Genes_CoVariate_Data[[geneName]] <- as.numeric(Genes_CoVariate_Data[[geneName]])

  # Plot histogram of gene expression data
  histograms <- hist(Genes_CoVariate_Data[[geneName]],
      main = paste(geneName, 'Gene Expression Data'), xlab = 'Gene Expression', col = "light blue")
  print(histograms)
  # Replace age values greater than 89 with 90
  Genes_CoVariate_Data[[Cont]][Genes_CoVariate_Data[[Cont]] == ">89"] <- 90
  # Convert continuous column to numeric
  Genes_CoVariate_Data[[Cont]] <- as.numeric(Genes_CoVariate_Data[[Cont]])

  # Create age groups
  Genes_CoVariate_Data$AgeGroup <- cut(Genes_CoVariate_Data[[Cont]],
                                       breaks = c(0, 30, 40, 50, 60, 70, 80, 90),
                                       labels = c('Under 30', '30-40', '40-50', '50-60', '60-70', '70-8
```

```
  # Plot scatter plot of gene expression by age
  Scatterplot <- ggplot(Genes_CoVariate_Data, aes(x = !!sym(Cont), y = !!sym(geneName), color = AgeGrou
    geom_point() +
    labs(title = paste('Gene Expression of', geneName, 'by', Cont),
         x = 'Age (yrs)',
         y = paste(geneName, 'Gene Expression'),
         color = "Age Group") +
    theme_classic()
  print(Scatterplot)

  # remove unknown values from data set
  Genes_CoVariate_Data[Genes_CoVariate_Data == "unknown"] <- NA
  Genes_CoVariate_Data<- na.omit(Genes_CoVariate_Data)

# add color pallete to color boxplots
  colorPalette = c('light blue',' maroon')

# create box plots to show Gene Expression of ABCA3 by Sex and ICU Status
  boxPlots <- ggplot(Genes_CoVariate_Data, aes(x = !!sym(Cat1), y=!!sym(geneName), fill= !!sym(Cat2)) )
    # Add box plot
    geom_boxplot() +
    # add colors
    scale_fill_manual(values= colorPalette) +
    # change x value labels
    scale_x_discrete(labels = c("Female", "Male"))+
    # Change axis labels
    labs(title =paste('Gene Expression of',geneName,'by', Cat1,'and', Cat2), x = Cat1,y = 'Gene Expressi
    theme_classic()
  print(boxPlots)

}

#
GeneNames <- c('ABCA3')
#run Gene Plots function
GenePlots(MergedGenes, GeneNames, 'age', 'sex', 'icu_status')
```

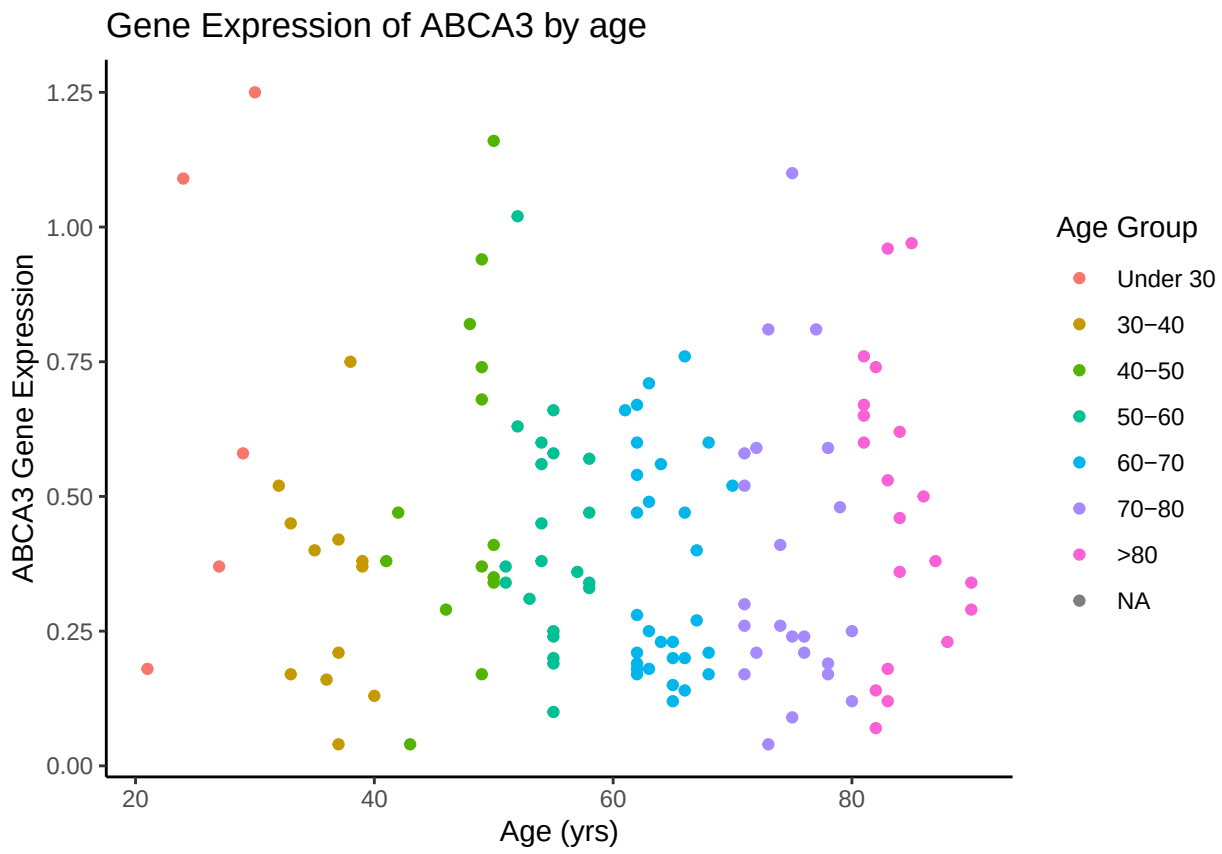
```
## $breaks
## [1] 0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
##
## $counts
## [1] 30 40 31 14  6  4  1
##
## $density
## [1] 1.19047619 1.58730159 1.23015873 0.55555556 0.23809524 0.15873016 0.03968254
##
## $mids
## [1] 0.1 0.3 0.5 0.7 0.9 1.1 1.3
##
## $xname
## [1] "Genes_CoVariate_Data[[geneName]]"
##
## $equidist
```

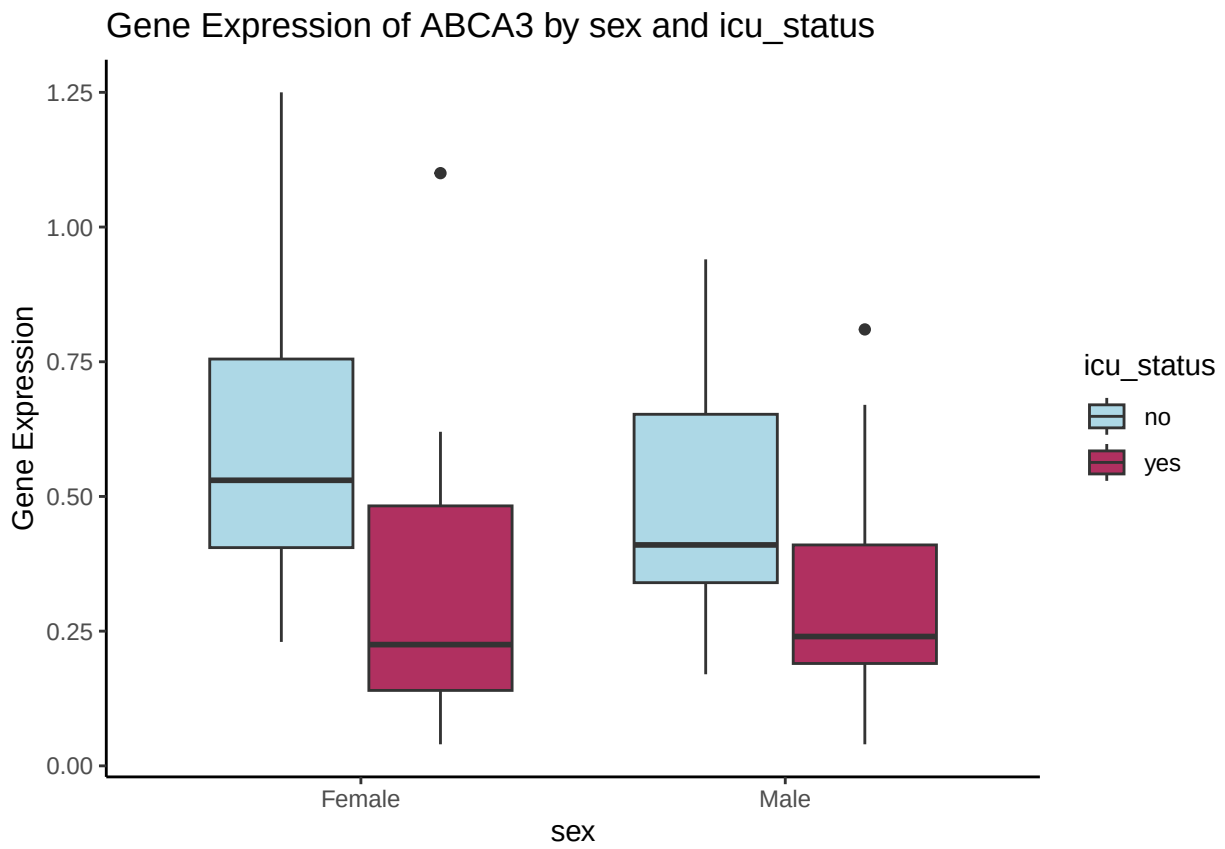
```
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
## Warning in GenePlots(MergedGenes, GeneNames, "age", "sex", "icu_status"): NAs
## introduced by coercion
```




```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```


Gene Expression of ABCA3 by age
ABCA3 Gene Expression
1.25
1.00
0.75
0.50
0.25
0.00
20
40
60
80
Age (yrs)
Age Group
Under 30
30−40
40−50
50−60
60−70
70−80
>80
NA



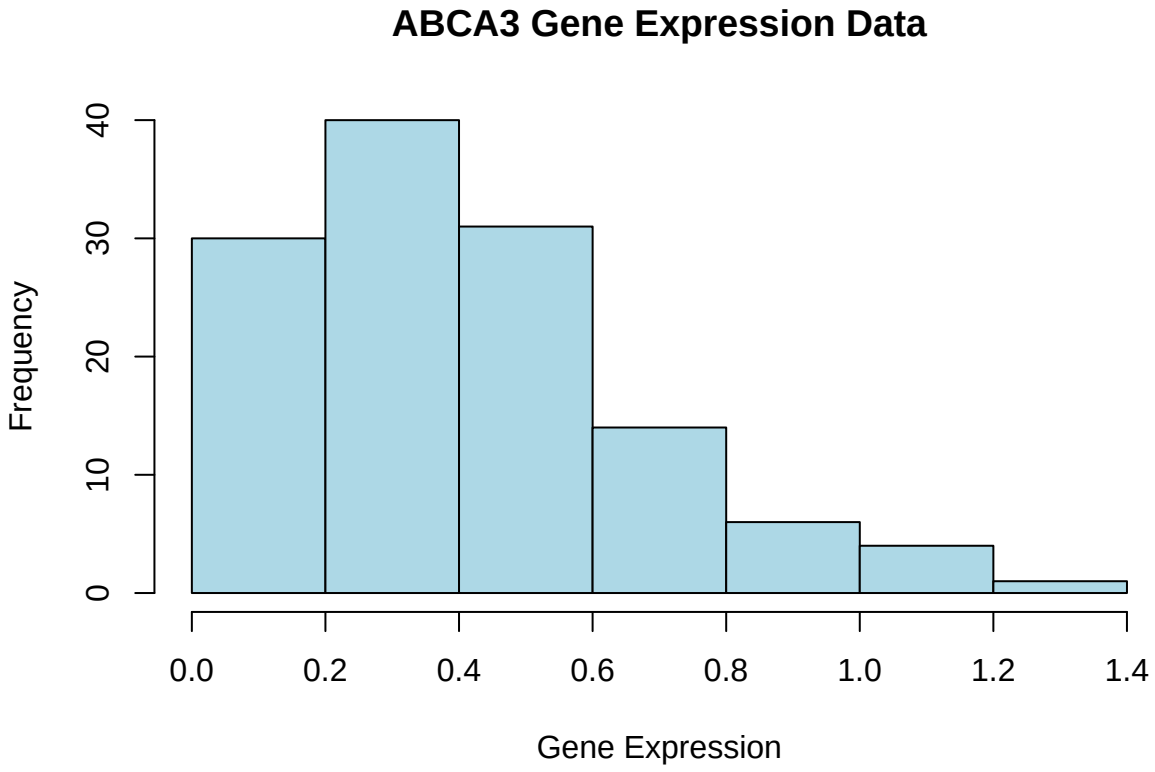
#Select 2 additional genes (for a total of 3 genes) to look at and implement a loop to generate your figures using the function you created (10 pts)

```
GeneNames <- c('ABCA3','ABHD1','AASS')

for(i in 1:length(GeneNames)) {
  GenePlots(MergedGenes, GeneNames[i], 'age', 'sex', 'icu_status')
}
```

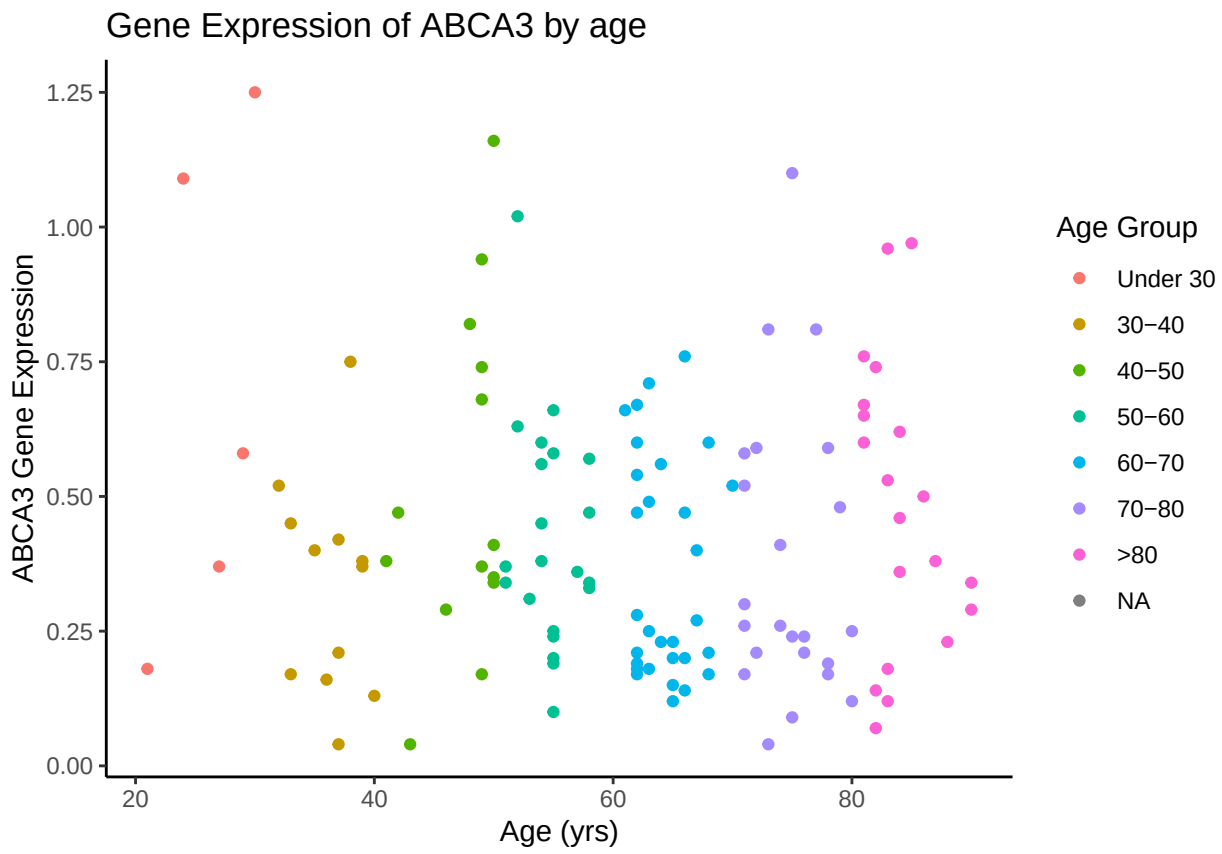
```
## $breaks
## [1] 0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
##
## $counts
## [1] 30 40 31 14  6  4  1
##
## $density
## [1] 1.19047619 1.58730159 1.23015873 0.55555556 0.23809524 0.15873016 0.03968254
##
## $mids
## [1] 0.1 0.3 0.5 0.7 0.9 1.1 1.3
##
## $xname
## [1] "Genes_CoVariate_Data[[geneName]]"
##
## $equidist
## [1] TRUE
##
```

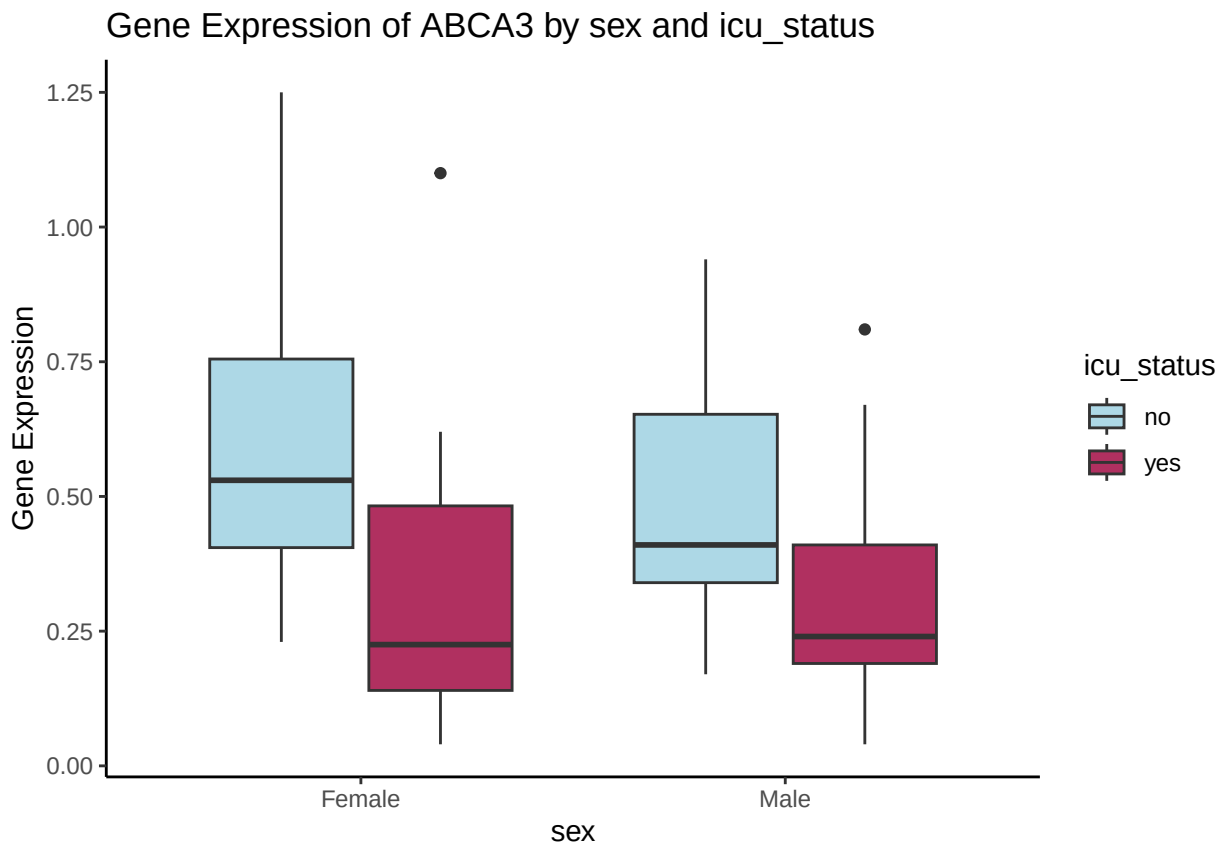
```
## attr(,"class")
## [1] "histogram"
```

```
## Warning in GenePlots(MergedGenes, GeneNames[i], "age", "sex", "icu_status"):
## NAs introduced by coercion
```

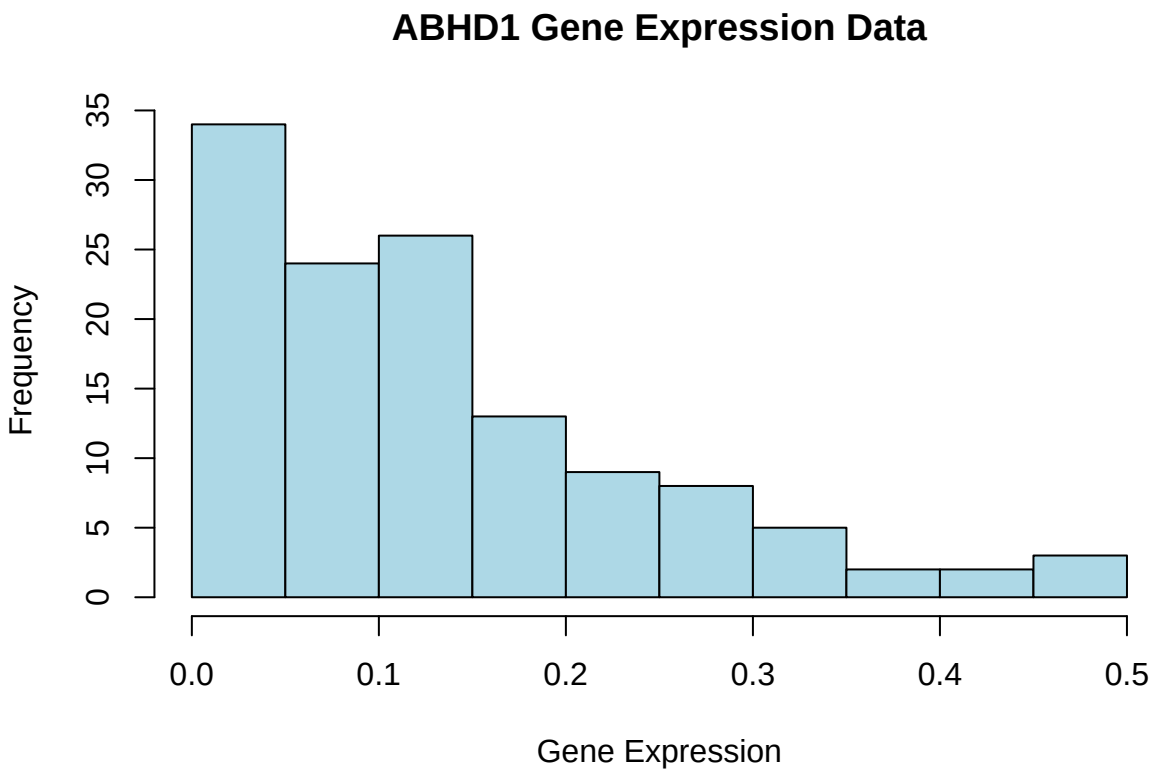



```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```

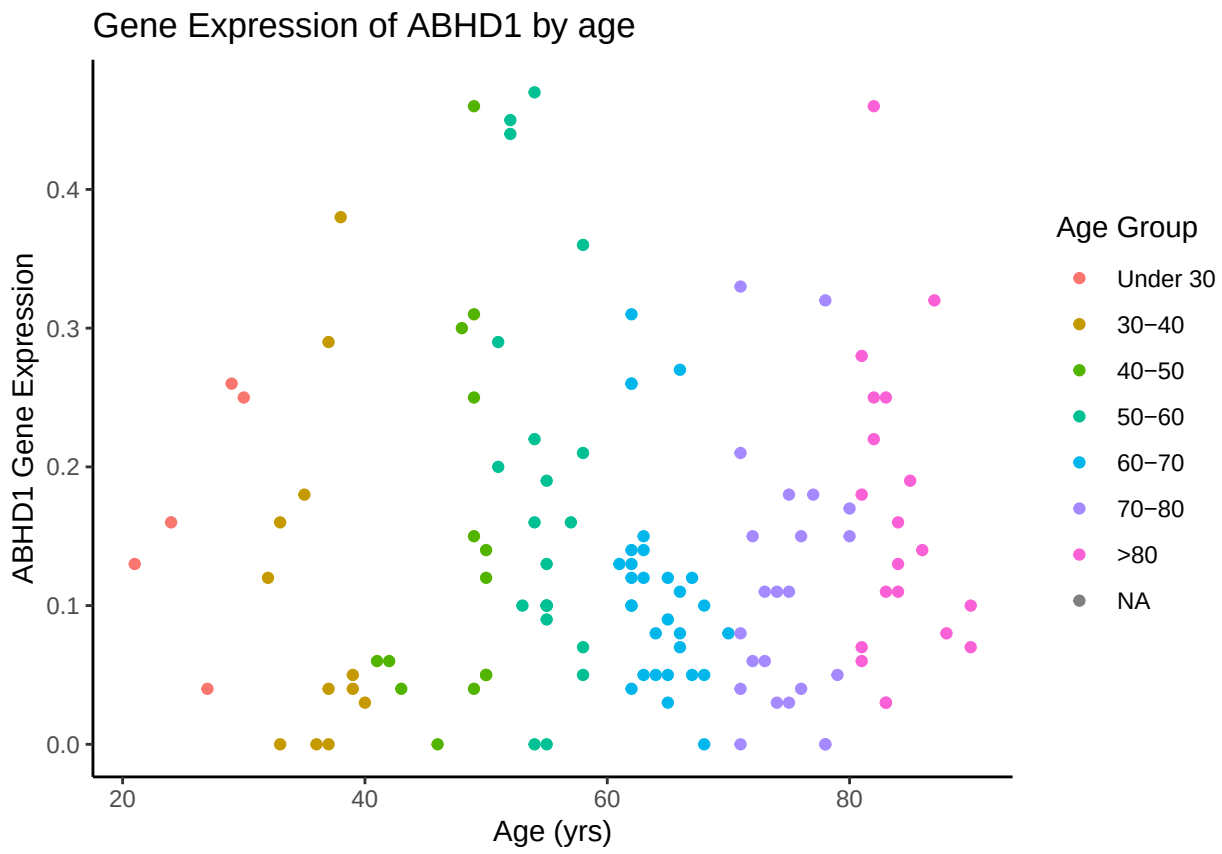

Gene Expression of ABCA3 by age
ABCA3 Gene Expression
1.25
1.00
0.75
0.50
0.25
0.00
20
40
60
80
Age (yrs)
Age Group
Under 30
30−40
40−50
50−60
60−70
70−80
>80
NA

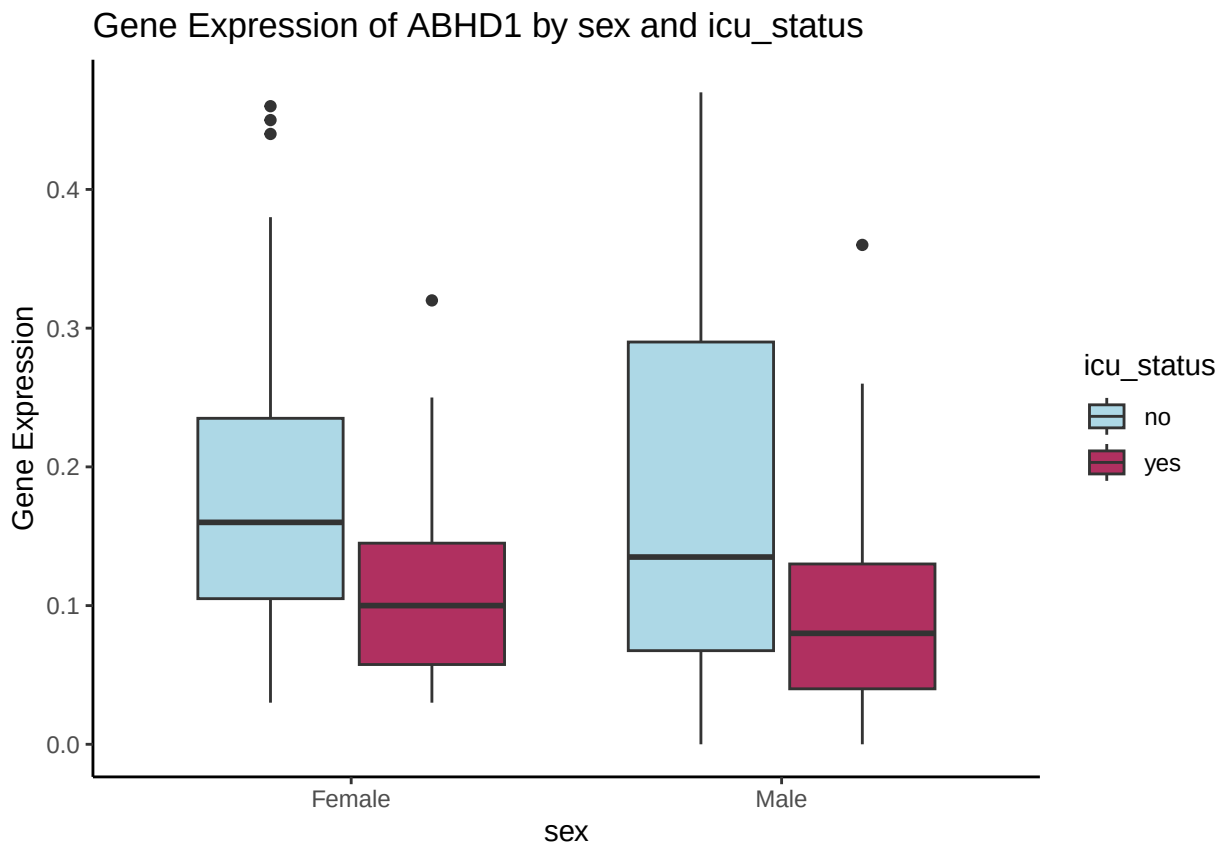



```
## $breaks
##  [1] 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50
##
## $counts
##  [1] 34 24 26 13  9  8  5  2  2  3
##
## $density
##  [1] 5.3968254 3.8095238 4.1269841 2.0634921 1.4285714 1.2698413 0.7936508
##  [8] 0.3174603 0.3174603 0.4761905
##
## $mids
##  [1] 0.025 0.075 0.125 0.175 0.225 0.275 0.325 0.375 0.425 0.475
##
## $xname
## [1] "Genes_CoVariate_Data[[geneName]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"

## Warning in GenePlots(MergedGenes, GeneNames[i], "age", "sex", "icu_status"):
## NAs introduced by coercion
```

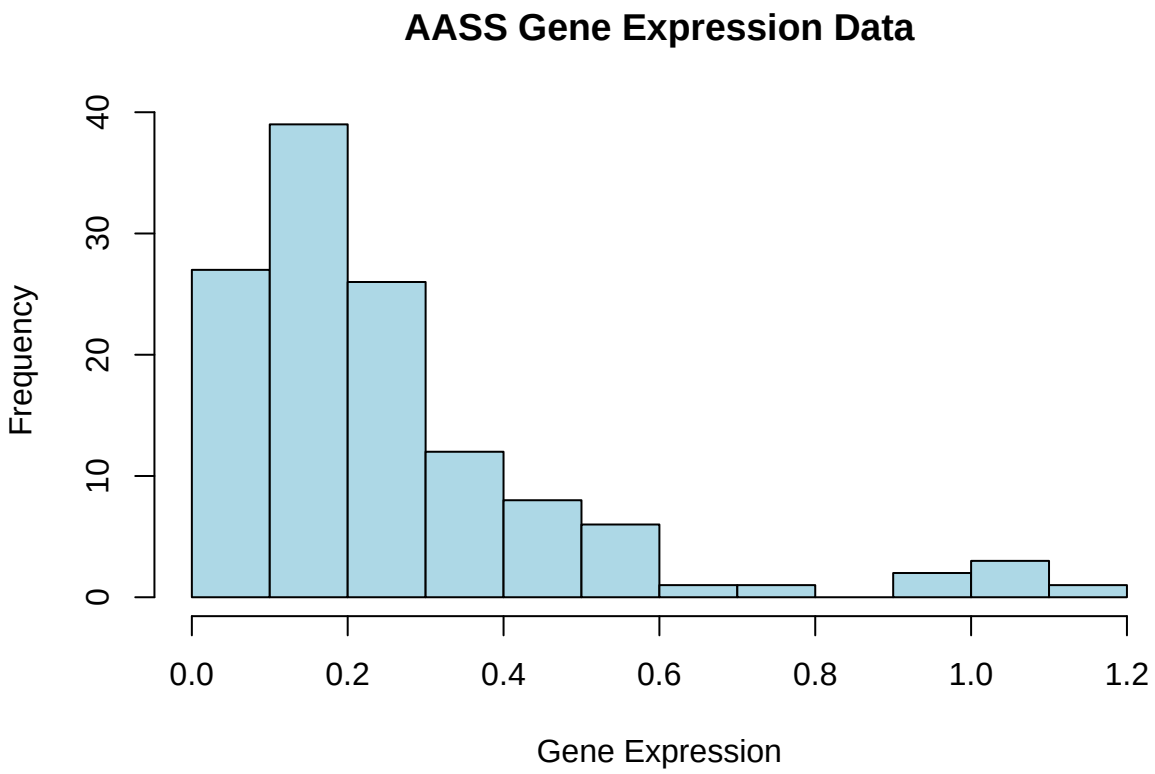



```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```

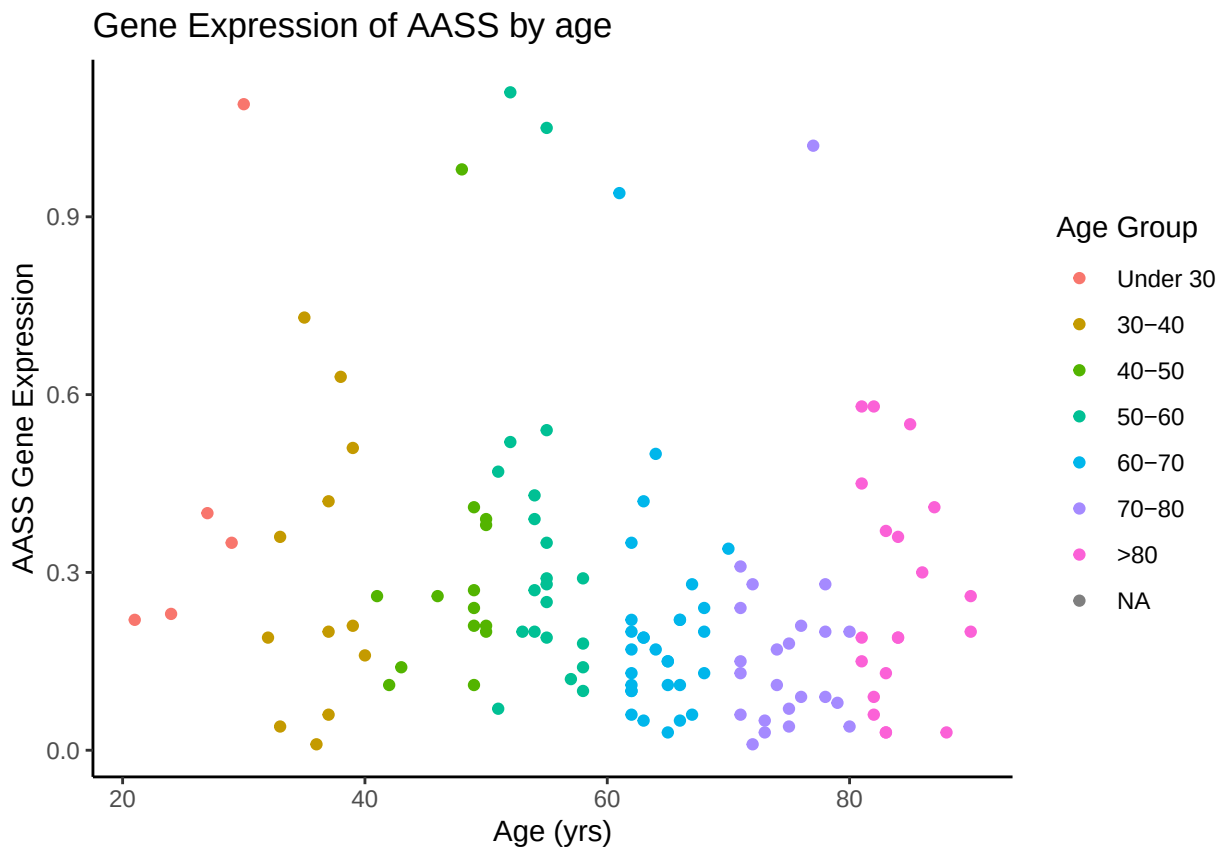

Gene Expression of ABHD1 by age
ABHD1 Gene Expression
0.0
0.1
0.2
0.3
0.4
20
40
60
80
Age (yrs)
Age Group
Under 30
30−40
40−50
50−60
60−70
70−80
>80
NA

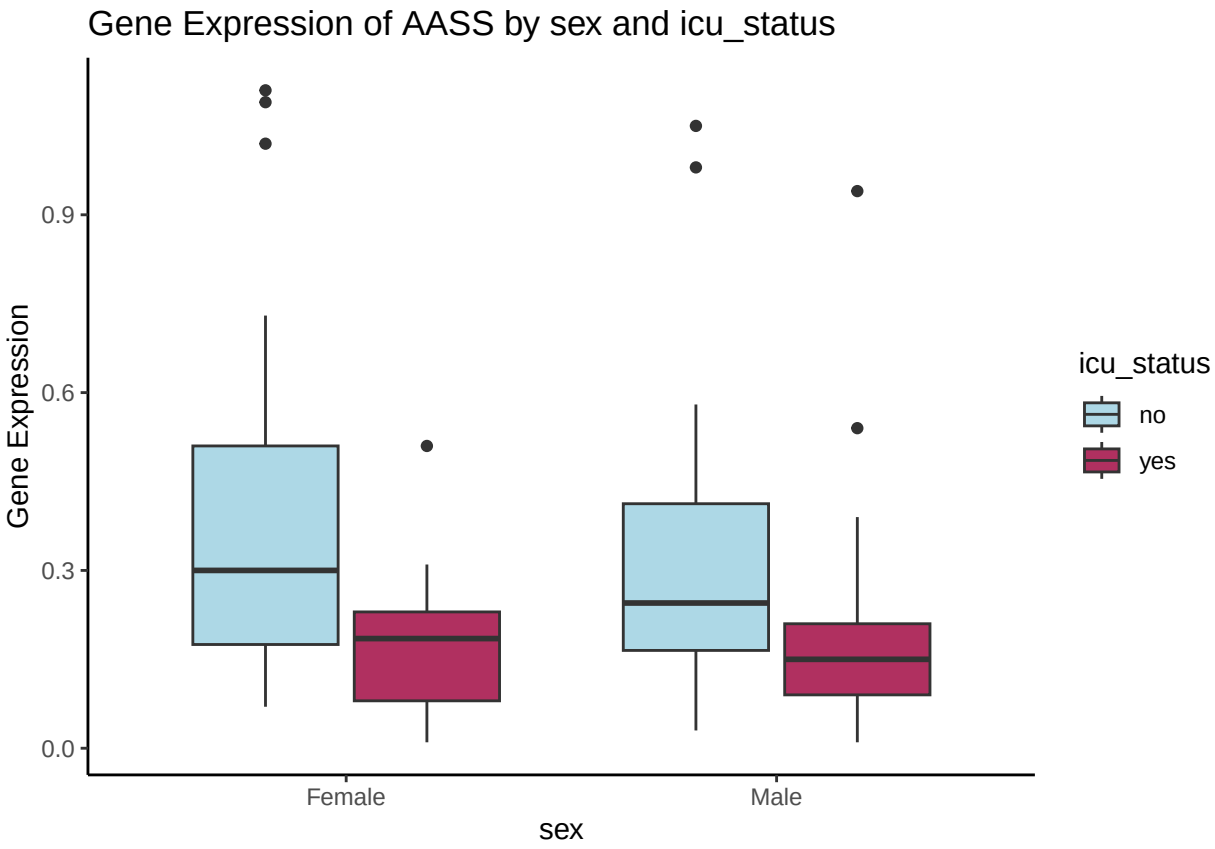



```
## $breaks
##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2
##
## $counts
##  [1] 27 39 26 12  8  6  1  1  0  2  3  1
##
## $density
##  [1] 2.14285714 3.09523810 2.06349206 0.95238095 0.63492063 0.47619048
##  [7] 0.07936508 0.07936508 0.00000000 0.15873016 0.23809524 0.07936508
##
## $mids
##  [1] 0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95 1.05 1.15
##
## $xname
## [1] "Genes_CoVariate_Data[[geneName]]"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"

## Warning in GenePlots(MergedGenes, GeneNames[i], "age", "sex", "icu_status"):
## NAs introduced by coercion
```




```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```


Gene Expression of AASS by age
AASS Gene Expression
0.0
0.3
0.6
0.9
20
40
60
80
Age (yrs)
Age Group
Under 30
30−40
40−50
50−60
60−70
70−80
>80
NA


Gene Expression of AASS by sex and icu_status
Gene Expression
0.0
0.3
0.6
0.9
Female
Male
sex
icu_status
no
yes