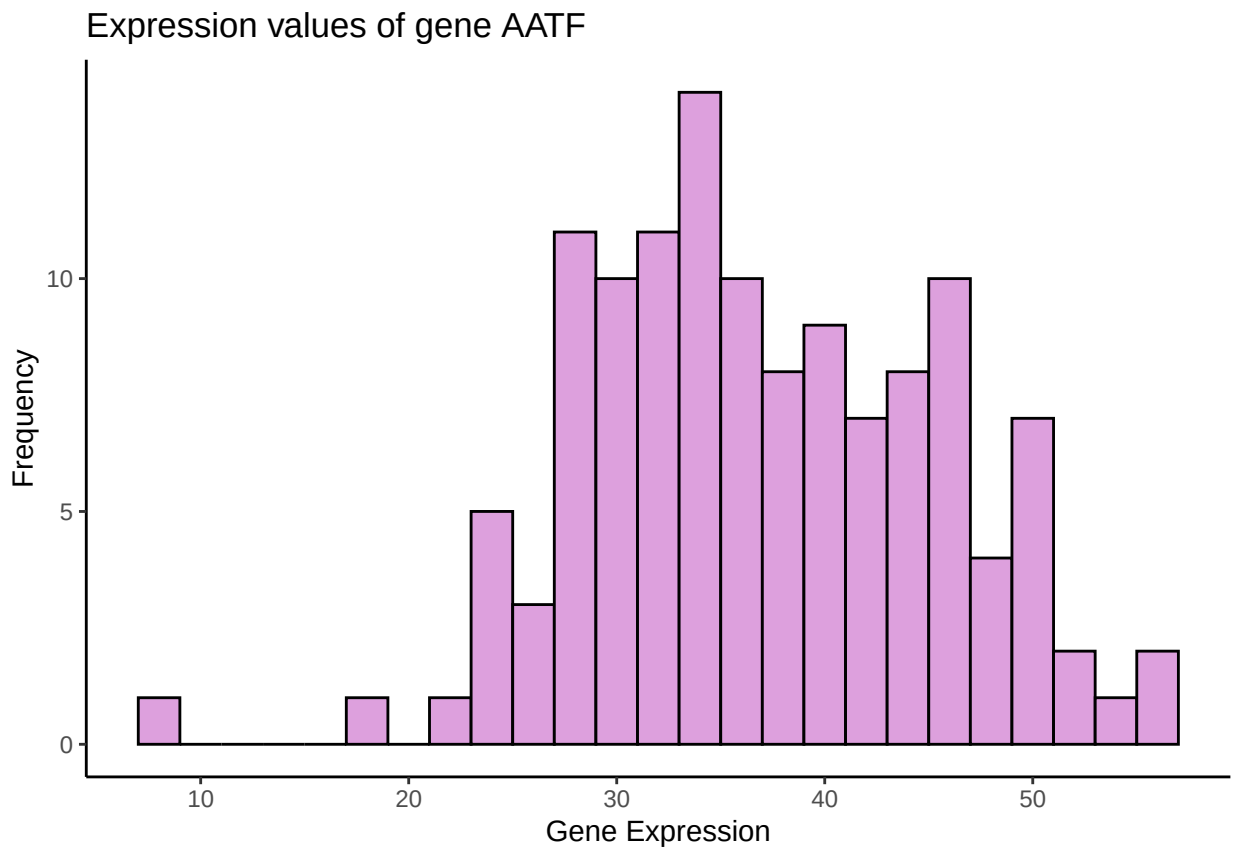
Final project 1

2025-07-12

```
long_gene <- genedata %>%
  pivot_longer(
    cols = -Gene,       # all columns except 'Gene' pivoted
    names_to = "participant_id", #naming column 2 according to the naming convention of my metadata in
#the two later
    values_to = "gene_expression"
  )
#print(long_gene)

#linking the two datasets
combined_data <- merge(long_gene, metadata, by = "participant_id")
#tail(combined_data)

#using a pipe to filter and select the data i want for my gene of interest AATF
AATFData <- combined_data %>%
  dplyr::filter(Gene == "AATF") %>%
  dplyr::select(participant_id, 'gene_expression', age, sex, icu_status) %>%
  dplyr::mutate(ICUStatus = ifelse(trimws(tolower(icu_status)) == 'yes', TRUE, FALSE))

#print(AATFData)
```

```
#creating histogram using ggplot. source: https://www.geeksforgeeks.org/r-language/histogram-in-r-using
ggplot(AATFData,aes(x = gene_expression)) +
  geom_histogram(binwidth = 2, color = "black", fill= "plum") +
  labs(x = "Gene Expression", y = "Frequency") +
  ggtitle("Expression values of gene AATF") +
  #scale_x_continuous(breaks=seq(2, 30, by = 2) +
  theme_classic()
```



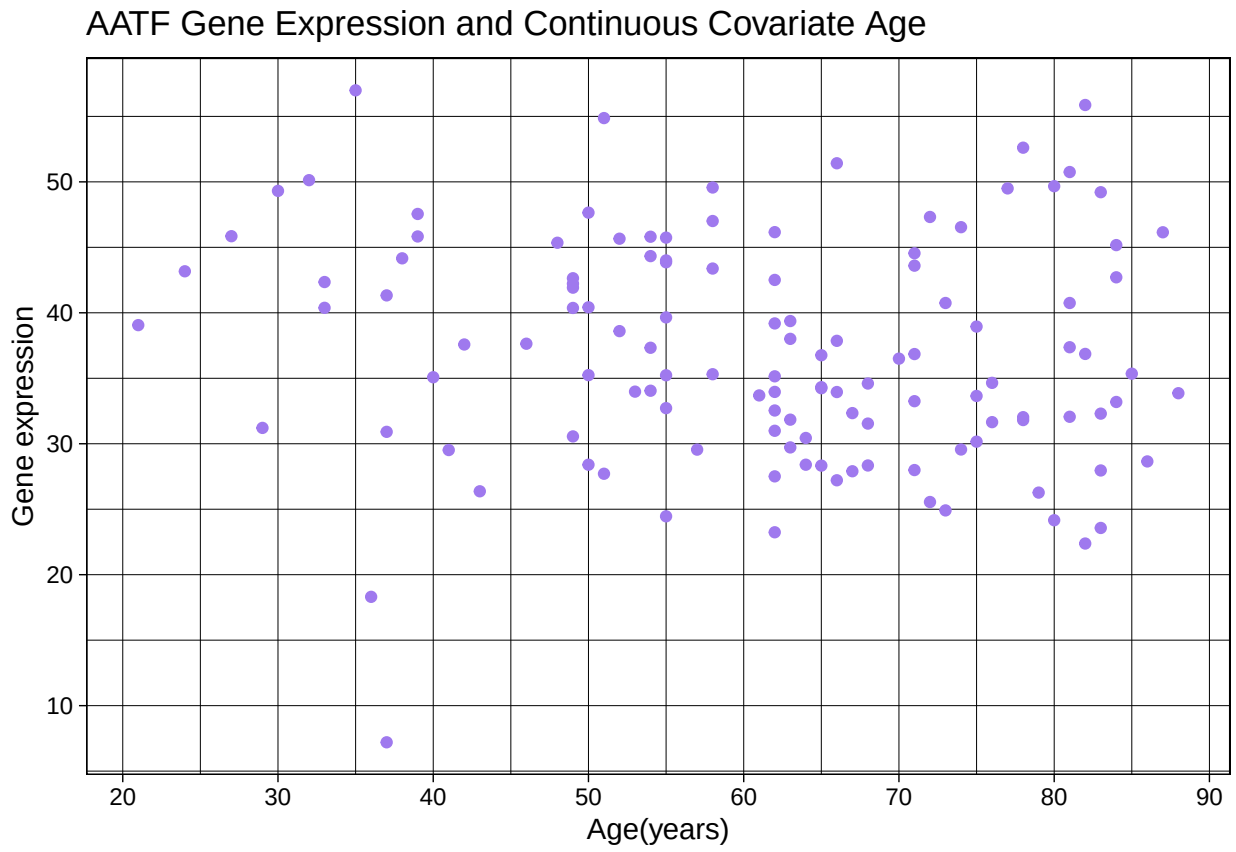
scatter plot

```
#colorPalette <- c('plum', 'mediumpurple2') #setting my colorpalette
AATFData$age <- as.numeric(AATFData$age) #converting my column age to numeric values to exclude NA valu
```

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

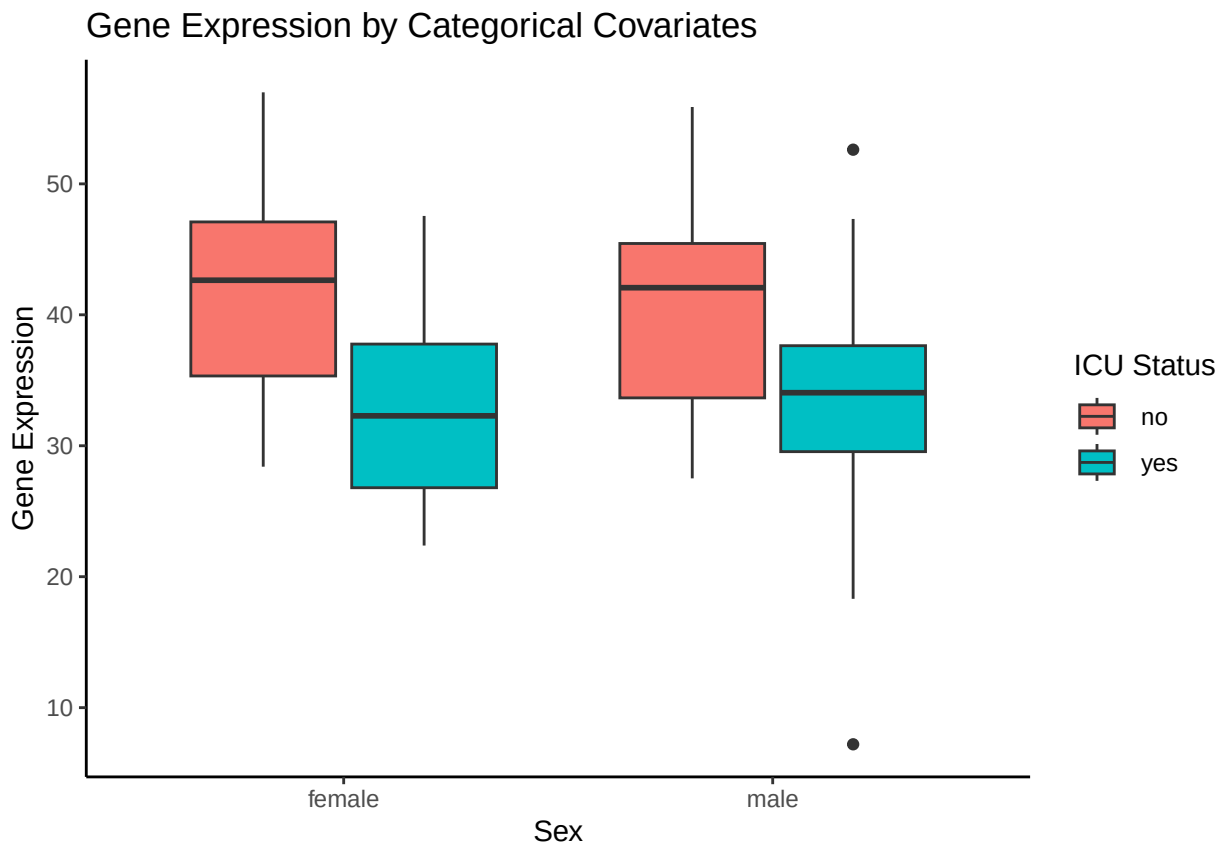
```
#and to ensure my interval breaks take effect properly. used chatgpt and https://vrcacademy.com/tutoria

#plotting scatterplot using ggplot function and set parameters
ggplot(AATFData, aes(x = age, y=gene_expression,)) +
    geom_point(color = 'mediumpurple2') +
  #when i first plotted without this function below, the age values were all over the place
  #this allows for better clarity and readability of the ages in intervals
  scale_x_continuous(breaks=seq(0, 100, by = 10)) +
  labs(title = "AATF Gene Expression and Continuous Covariate Age",
       x= 'Age(years)',
       y='Gene expression')+ #setting labels
  theme_linedraw()
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



Boxplot

```
#boxplot specifications plotting icu status, sex and gene expression

AATFData <- AATFData %>%
 mutate(sex_standard = str_trim(tolower(sex))) #standardized the format of sex column
AATFData_sex <- AATFData %>%
  filter(!sex_standard %in% c("unknown", "", "na", "n/a") & !is.na(sex_standard))
#chat gpt was used here to understand what mistake i was making when trying to filter out the unwanted

ggplot(AATFData_sex, aes(x = sex_standard, y = gene_expression, fill = icu_status)) +
    geom_boxplot()+
    theme_classic() +
    labs(title = "Gene Expression by Categorical Covariates",
         x = "Sex",
         y = "Gene Expression",
         fill = "ICU Status")
```

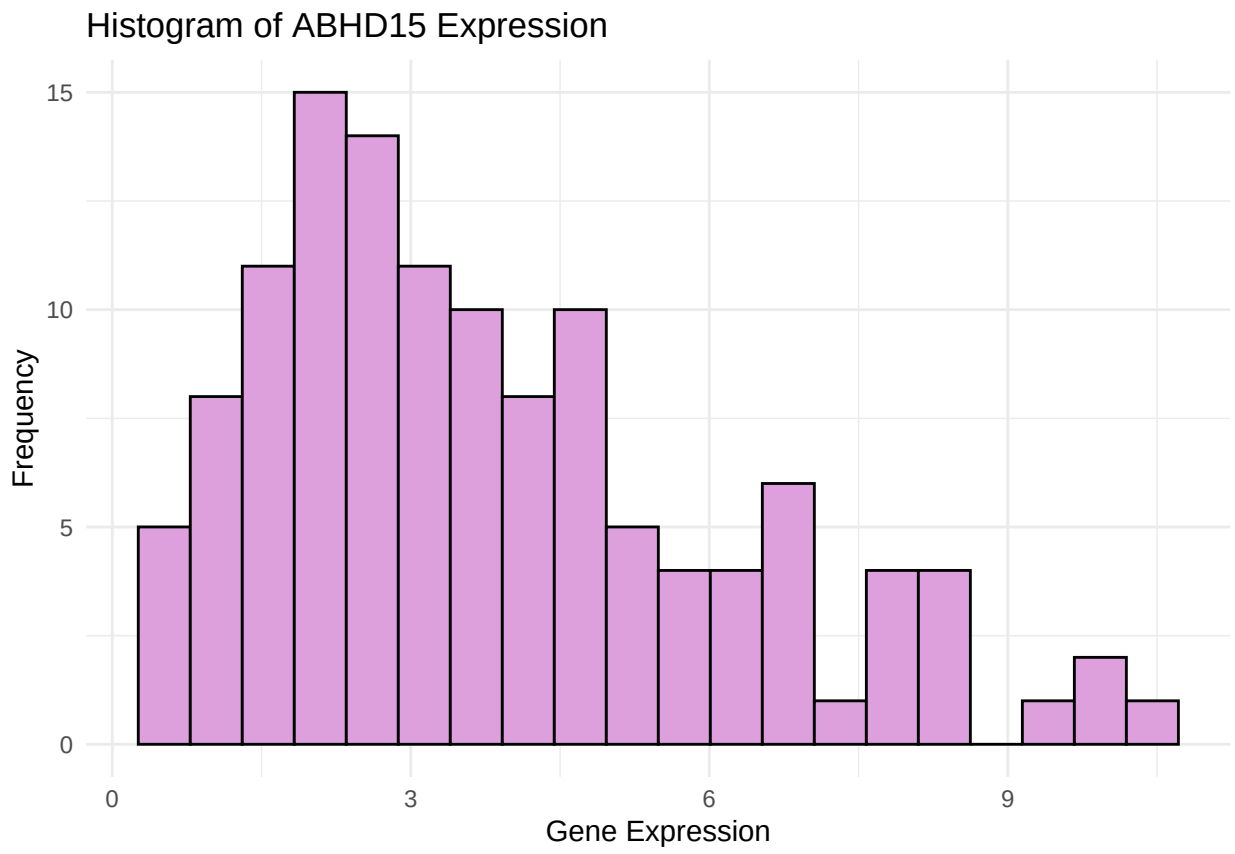



```
my_function <- function(data, gene_list, continuous_cov, categorical_cov1, categorical_cov2) {

  for (gene_name in gene_list) {

    # Setting filters and paramaters within the function
    gene_data <- data %>%
      filter(Gene == gene_name) %>%
      #!!sym is used here to tell r to take the string stored in the variable provided
      #and evaluate it as a column name
      select(participant_id, gene_expression, !!sym(continuous_cov),
             !!sym(categorical_cov1), !!sym(categorical_cov2)) %>%
      # ensuring continuous covariate is numeric
      mutate(!!sym(continuous_cov) := as.numeric(!!sym(continuous_cov))) %>%
      # Cleaning categorical variables
      mutate(across(c(!!sym(categorical_cov1), !!sym(categorical_cov2)), ~ str_trim(tolower(.)))) %>%
      #accessing columns for categorical variable and removing unwanted values
      filter(!is.na(!!sym(categorical_cov1)) & !(!!sym(categorical_cov1)
                                                 %in% c("unknown", "", "na", "n/a")))

    #i asked chatgpt here to evaluate my code and it was used to determine some corrections
    #the function parameters were updated to allow user input when using the function
    #this update allows for function usability across other data sets

    # Histogram
    histogram <- ggplot(gene_data, aes(x = gene_expression)) +
      geom_histogram(fill = "plum", color = "black", bins=20) +
```

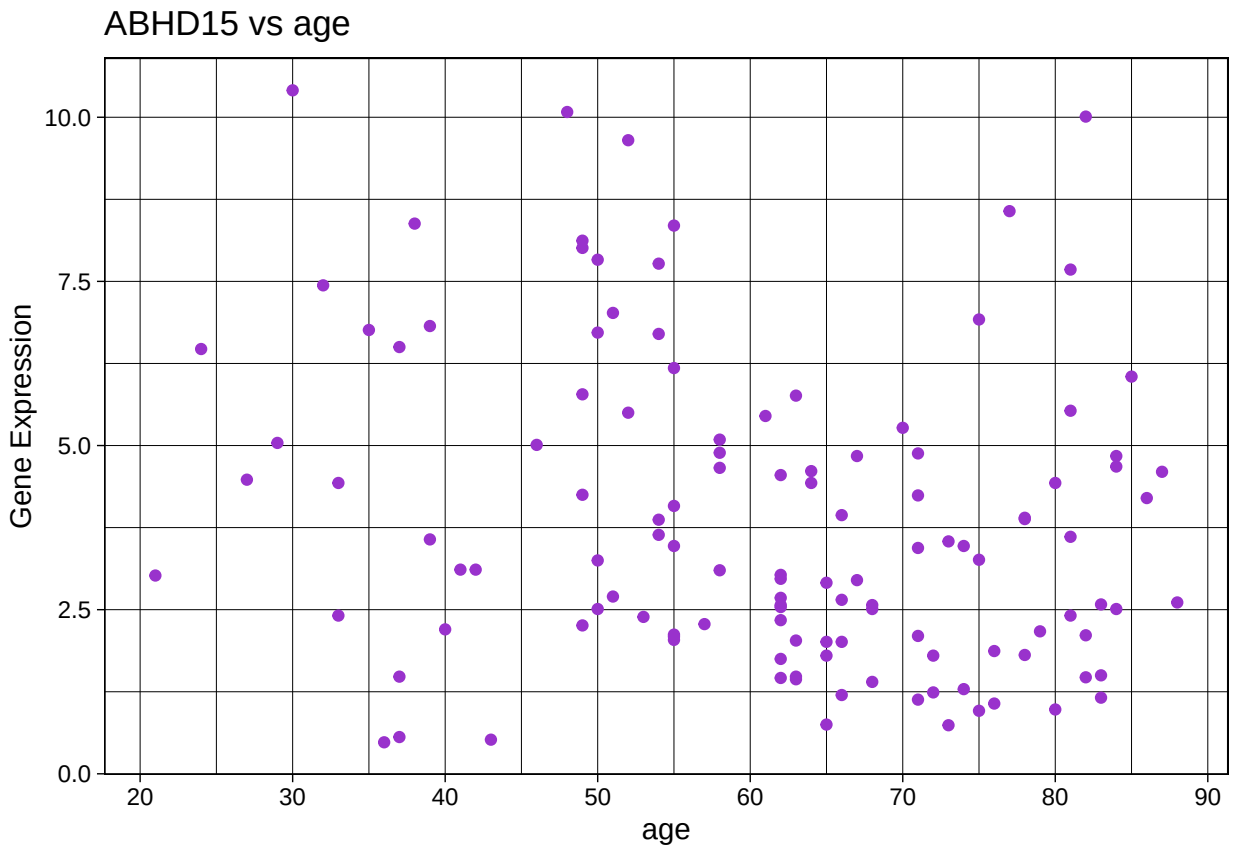
```
      labs(title = paste("Histogram of", gene_name, "Expression"),
           x = "Gene Expression", y = "Frequency") +
      theme_minimal()

    # Scatterplot
    scatterplot <- ggplot(gene_data, aes(x = !!sym(continuous_cov), y = gene_expression)) +
      geom_point(color = "darkorchid") +
      scale_x_continuous(breaks = seq(0, 100, by = 10)) +
      labs(title = paste(gene_name, "vs", continuous_cov),
           x = continuous_cov, y = "Gene Expression") +
      theme_linedraw()

    # Boxplot
    boxplot <- ggplot(gene_data, aes(x = !!sym(categorical_cov1), y = gene_expression,
                                     fill = !!sym(categorical_cov2))) +
      geom_boxplot() +
      scale_fill_manual(values=c("hotpink3","lightskyblue")) +
      labs(title = paste("Expression of", gene_name, "by", categorical_cov1),
           x = categorical_cov1, y = "Gene Expression", fill = categorical_cov2) +
      theme_classic()

    print(histogram)
    print(scatterplot)
    print(boxplot)
  }
}
```

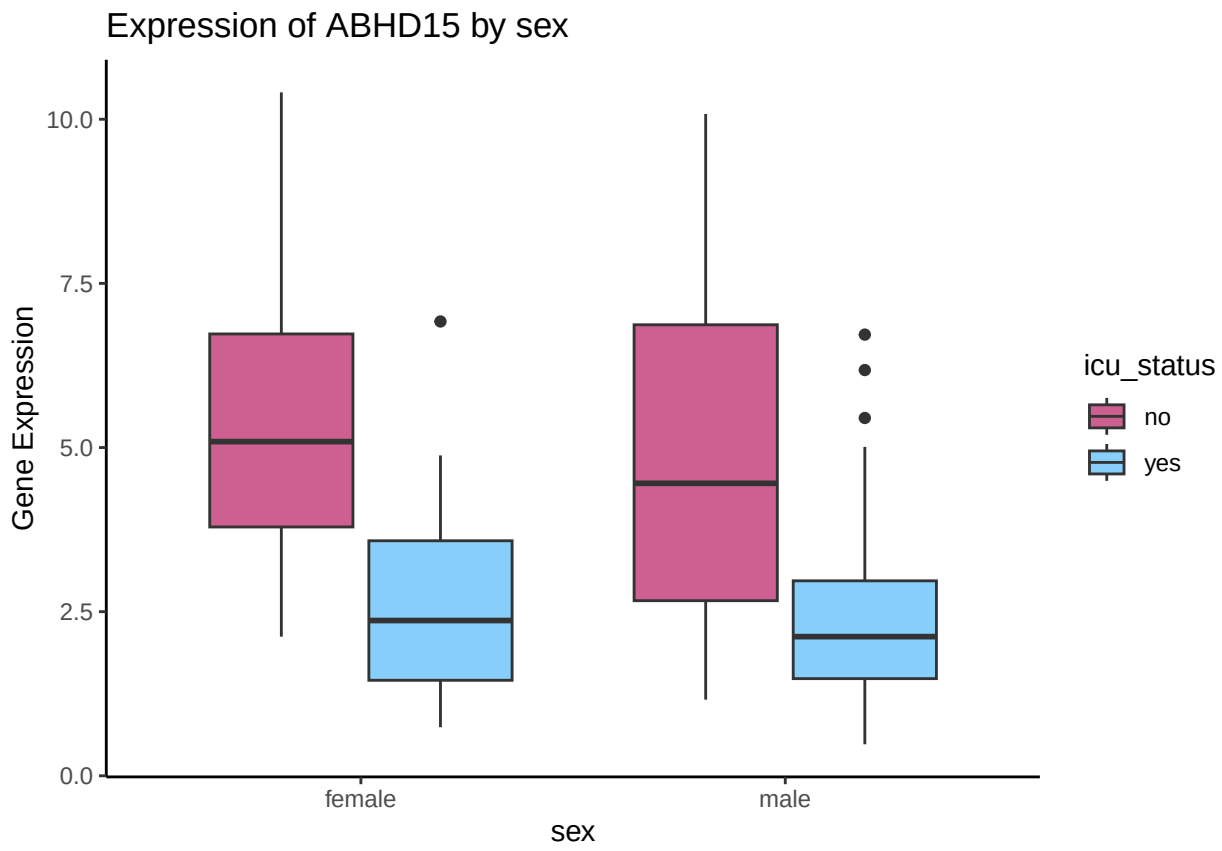
```
#calling my previously defined function and setting values for the parameters
my_function(data = combined_data,
            gene_list = c("ABHD15","ABI1","AATF"),
            continuous_cov = "age",
            categorical_cov1 = "sex",
            categorical_cov2 = "icu_status")
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `age = as.numeric(age)`.
## Caused by warning:
## ! NAs introduced by coercion
```

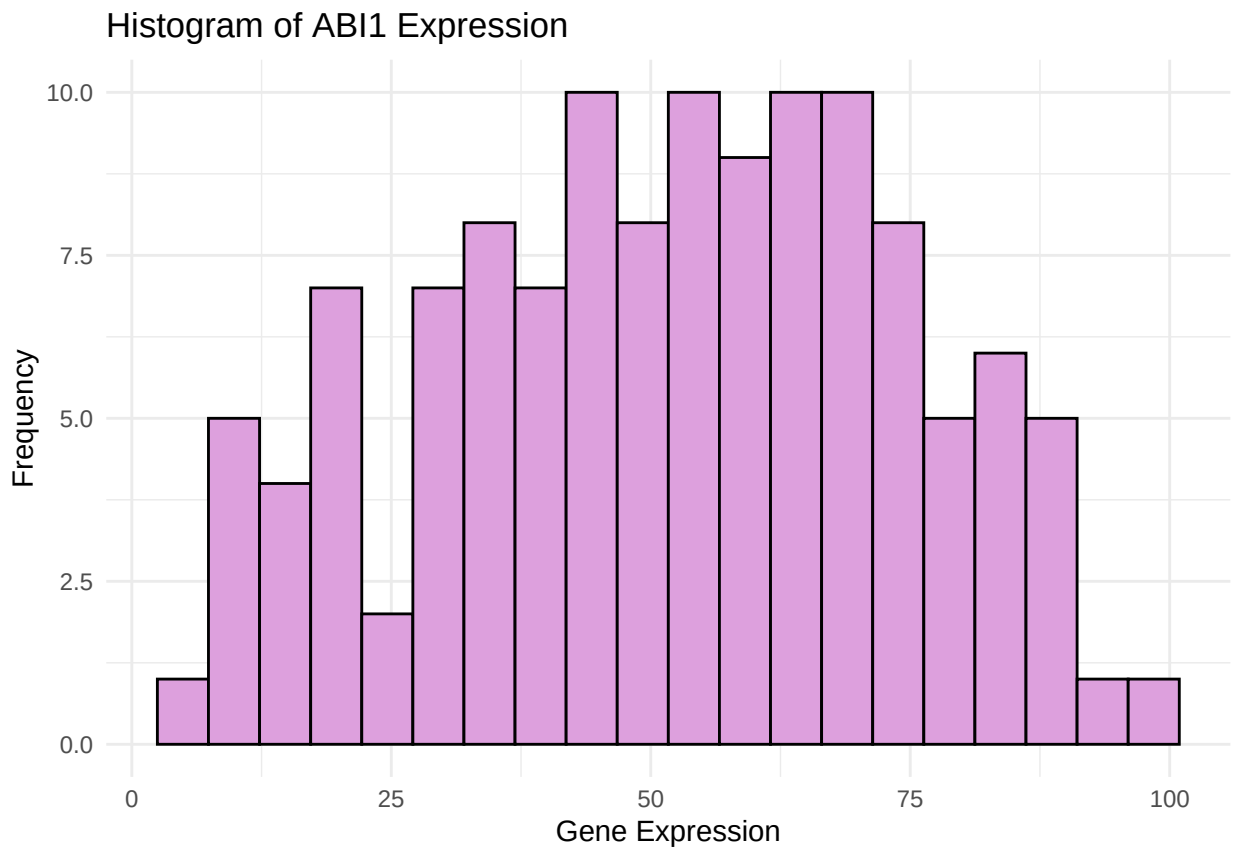



```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

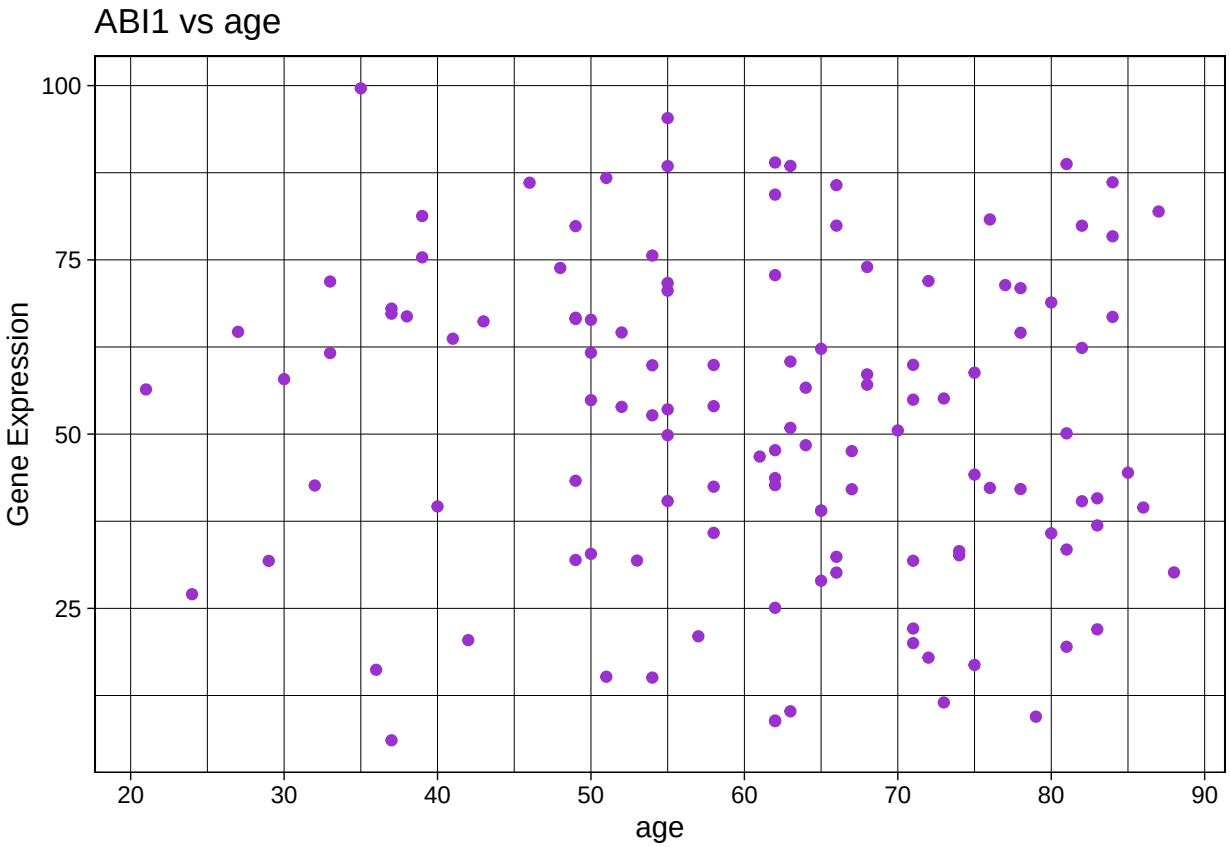

ABHD15 vs age
Gene Expression
10.0
7.5
5.0
2.5
0.0
20
30
40
50
60
70
80
90
age

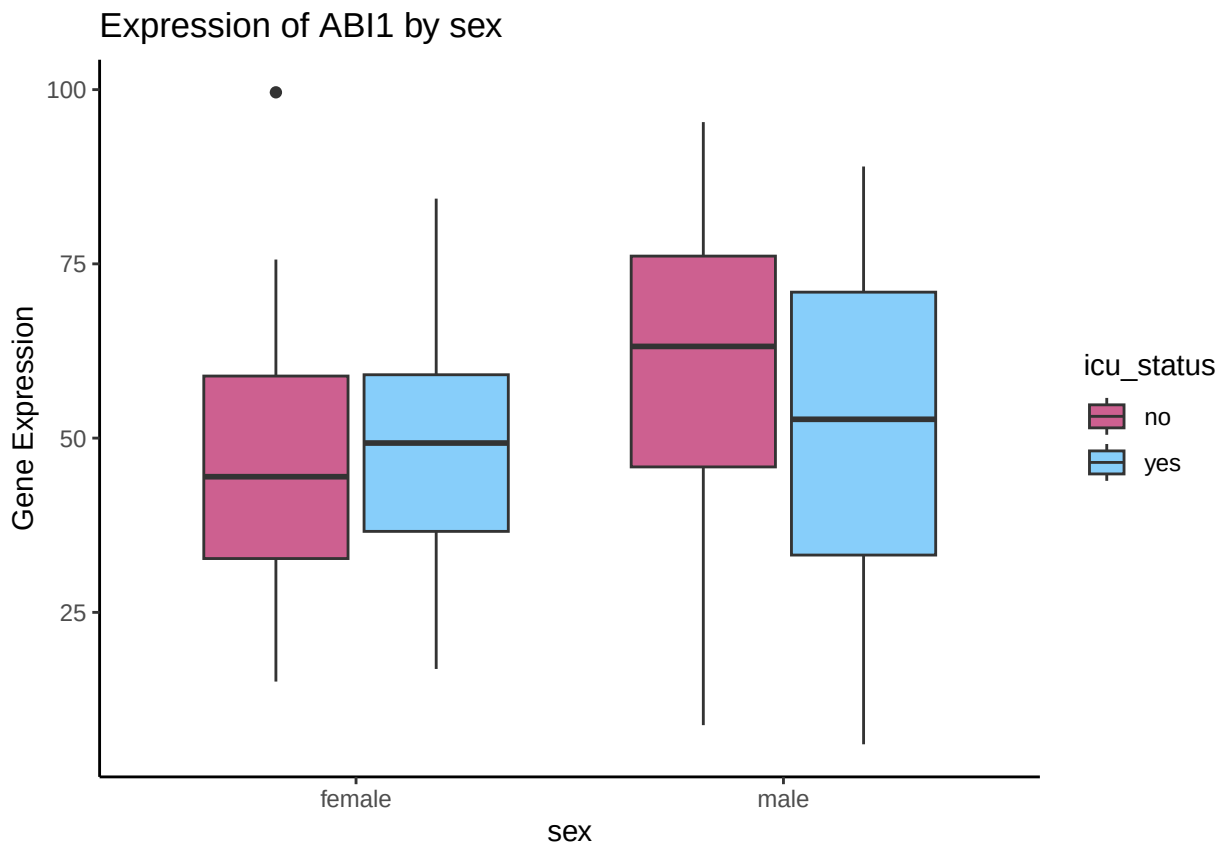



```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `age = as.numeric(age)`.
## Caused by warning:
## ! NAs introduced by coercion
```

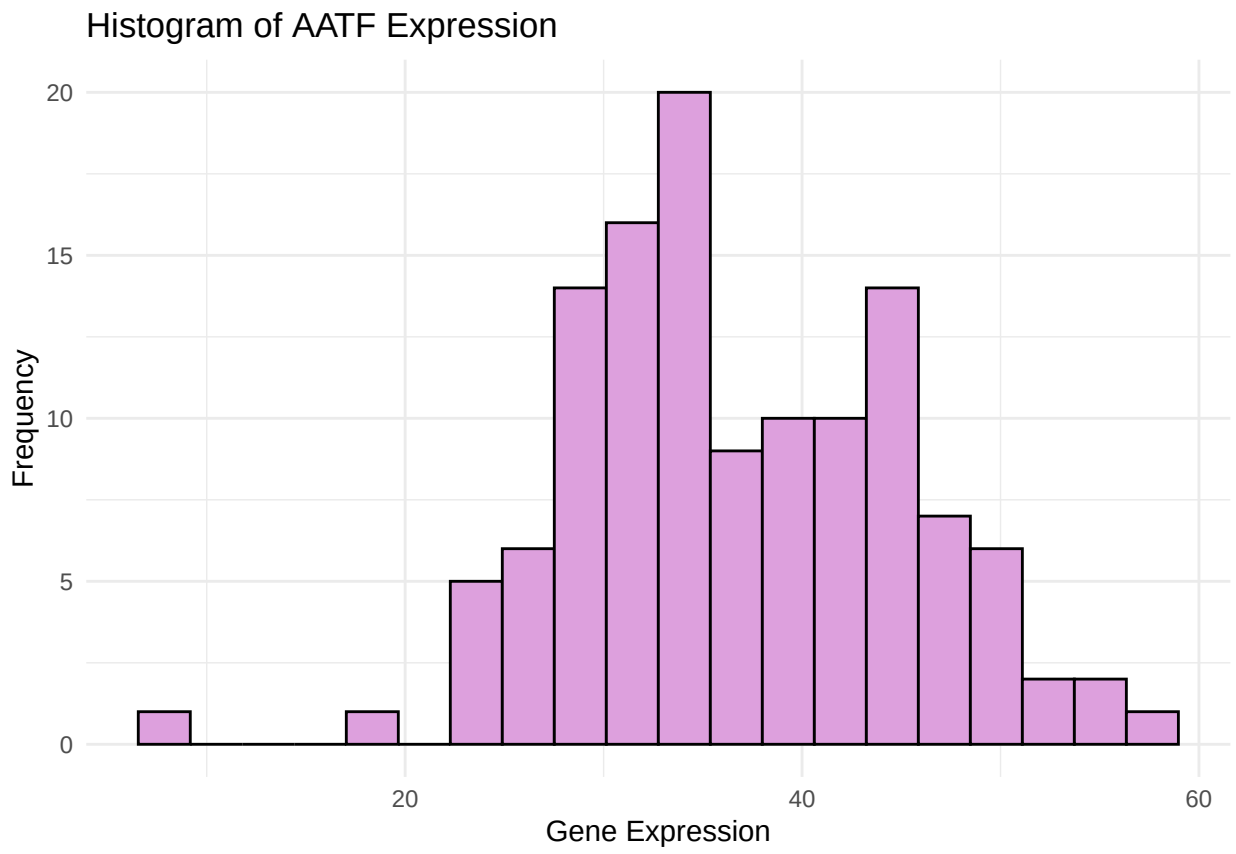



```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

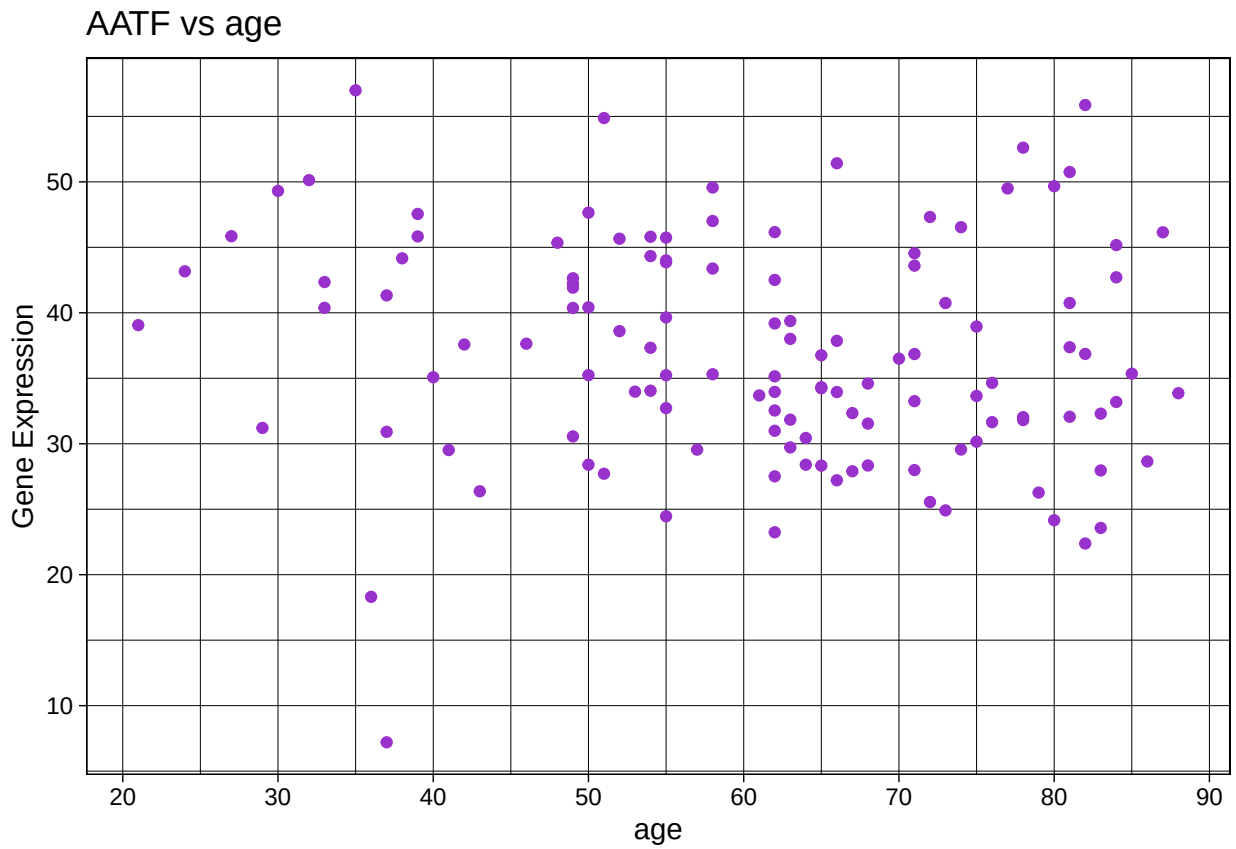

ABI1 vs age
Gene Expression
100
75
50
25
20
30
40
50
60
70
80
90
age

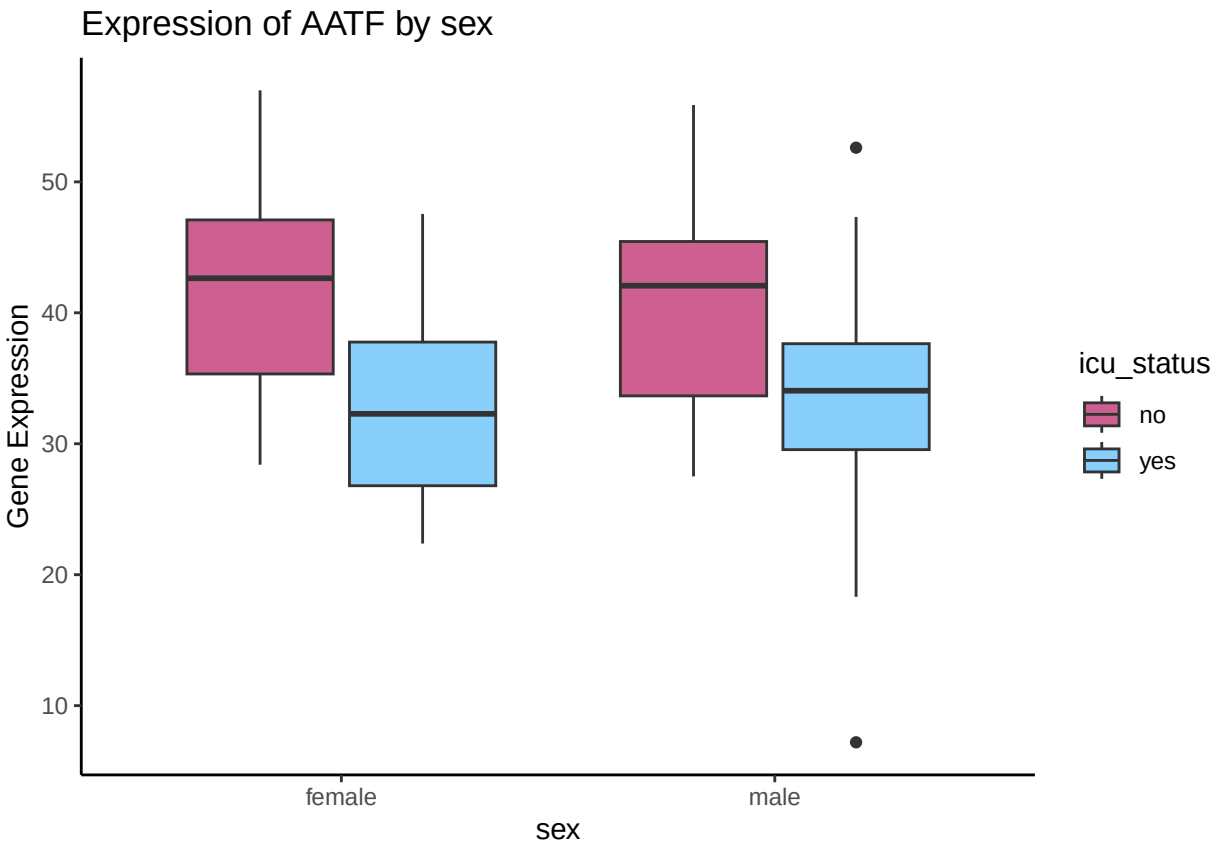



```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `age = as.numeric(age)`.
## Caused by warning:
## ! NAs introduced by coercion
```




```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```


AATF vs age
Gene Expression
age
10
20
30
40
50
20
30
40
50
60
70
80
90


Expression of AATF by sex
Gene Expression
50
40
30
20
10
female
male
sex
icu_status
no
yes