Final project 1

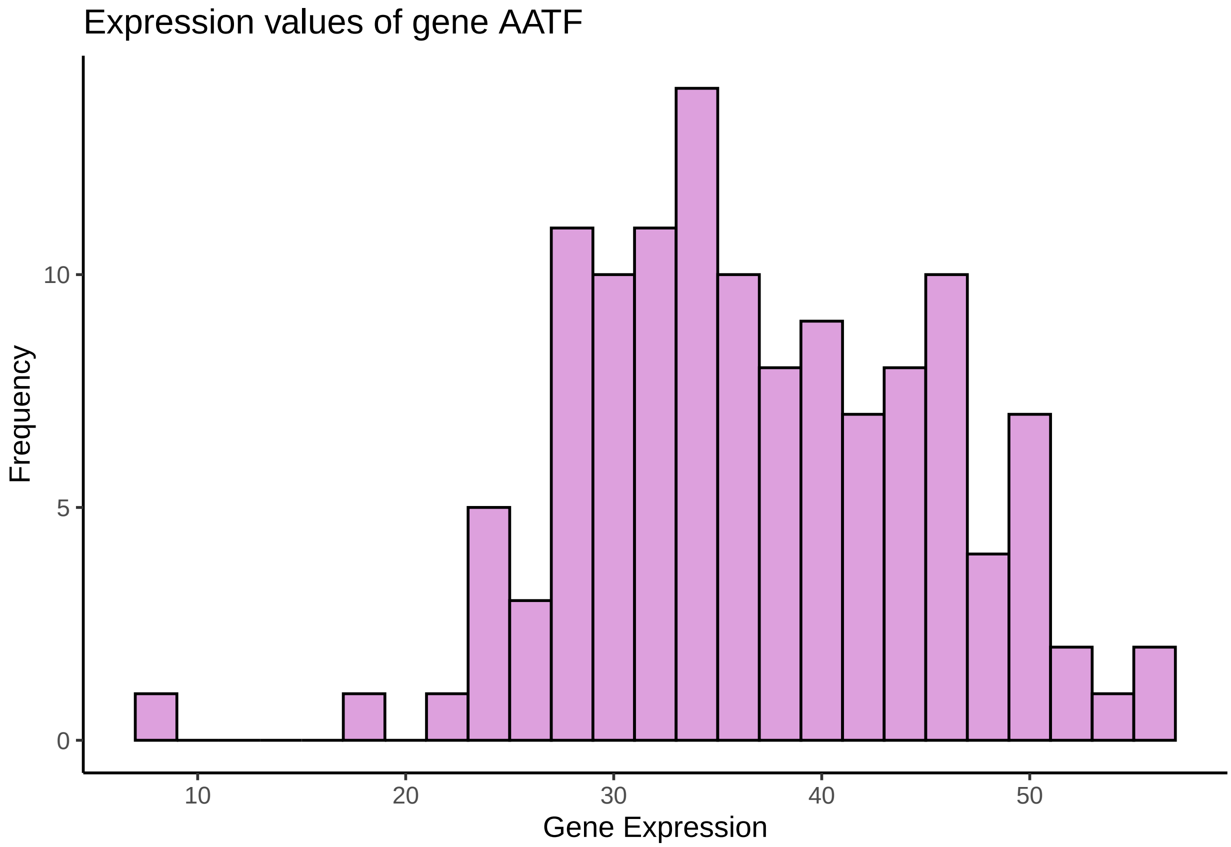
2025-07-12

```
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)
```

```
##                 participant_id  Gene gene_expression geo_accession
## 12495 NONCOVID_26_36y_male_ICU  ABI1           16.19    GSM4753146
## 12496 NONCOVID_26_36y_male_ICU  A1BG            0.22    GSM4753146
## 12497 NONCOVID_26_36y_male_ICU AADAC            0.00    GSM4753146
## 12498 NONCOVID_26_36y_male_ICU AANAT            0.00    GSM4753146
## 12499 NONCOVID_26_36y_male_ICU ABCG2            0.00    GSM4753146
## 12500 NONCOVID_26_36y_male_ICU  ABI2            0.32    GSM4753146
##                      status X.Sample_submission_date last_update_date type
## 12495 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 12496 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 12497 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 12498 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 12499 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
## 12500 Public on Aug 29 2020              Aug 28 2020      Aug 29 2020  SRA
##       channel_count             source_name_ch1 organism_ch1
## 12495             1 Leukocytes from whole blood Homo sapiens
## 12496             1 Leukocytes from whole blood Homo sapiens
## 12497             1 Leukocytes from whole blood Homo sapiens
## 12498             1 Leukocytes from whole blood Homo sapiens
## 12499             1 Leukocytes from whole blood Homo sapiens
## 12500             1 Leukocytes from whole blood Homo sapiens
##                    disease_status age  sex icu_status apacheii charlson_score
## 12495 disease state: non-COVID-19  36 male        yes       40              3
## 12496 disease state: non-COVID-19  36 male        yes       40              3
## 12497 disease state: non-COVID-19  36 male        yes       40              3
## 12498 disease state: non-COVID-19  36 male        yes       40              3
## 12499 disease state: non-COVID-19  36 male        yes       40              3
## 12500 disease state: non-COVID-19  36 male        yes       40              3
##       mechanical_ventilation ventilator.free_days
## 12495                    yes                    0
## 12496                    yes                    0
## 12497                    yes                    0
## 12498                    yes                    0
## 12499                    yes                    0
## 12500                    yes                    0
##       hospital.free_days_post_45_day_followup ferritin.ng.ml. crp.mg.l.
## 12495                                       0         unknown   unknown
## 12496                                       0         unknown   unknown
## 12497                                       0         unknown   unknown
## 12498                                       0         unknown   unknown
## 12499                                       0         unknown   unknown
## 12500                                       0         unknown   unknown
##       ddimer.mg.l_feu. procalcitonin.ng.ml.. lactate.mmol.l. fibrinogen sofa
## 12495          unknown               unknown         unknown    unknown   15
## 12496          unknown               unknown         unknown    unknown   15
## 12497          unknown               unknown         unknown    unknown   15
## 12498          unknown               unknown         unknown    unknown   15
## 12499          unknown               unknown         unknown    unknown   15
## 12500          unknown               unknown         unknown    unknown   15
```

```r
#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)
```

```r
#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()
```



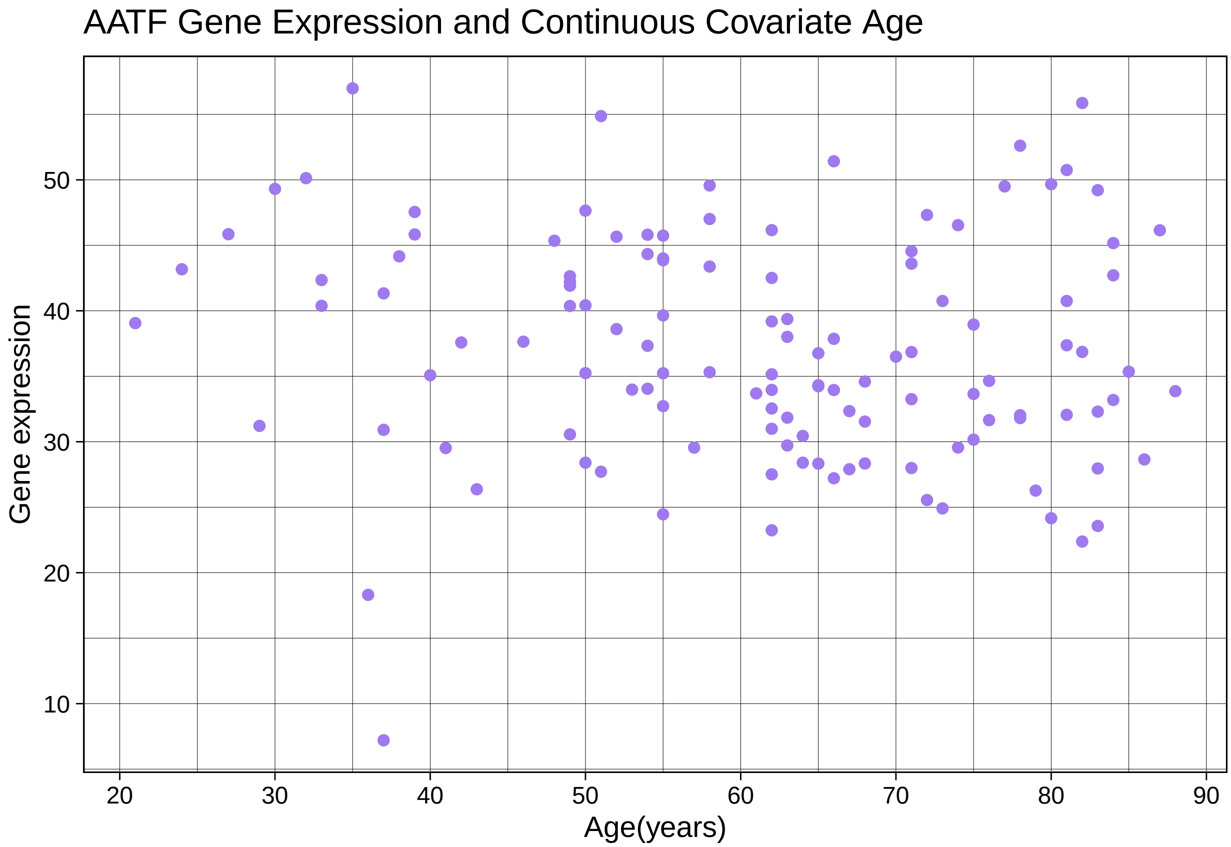
scatter plot

```
#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
```

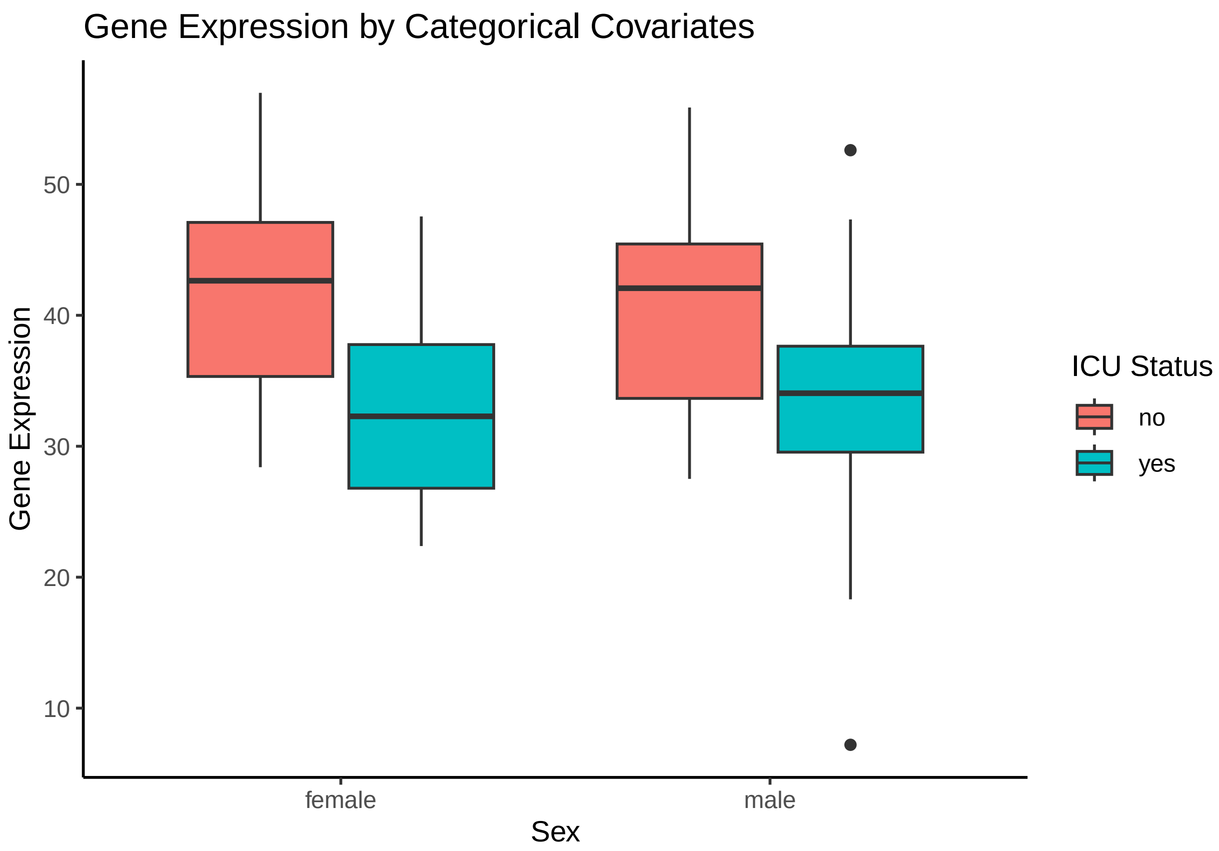
```
#and to ensure my interval breaks take effect properly. used chatgpt and https://vrcacademy.com/tutoria

#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")
```

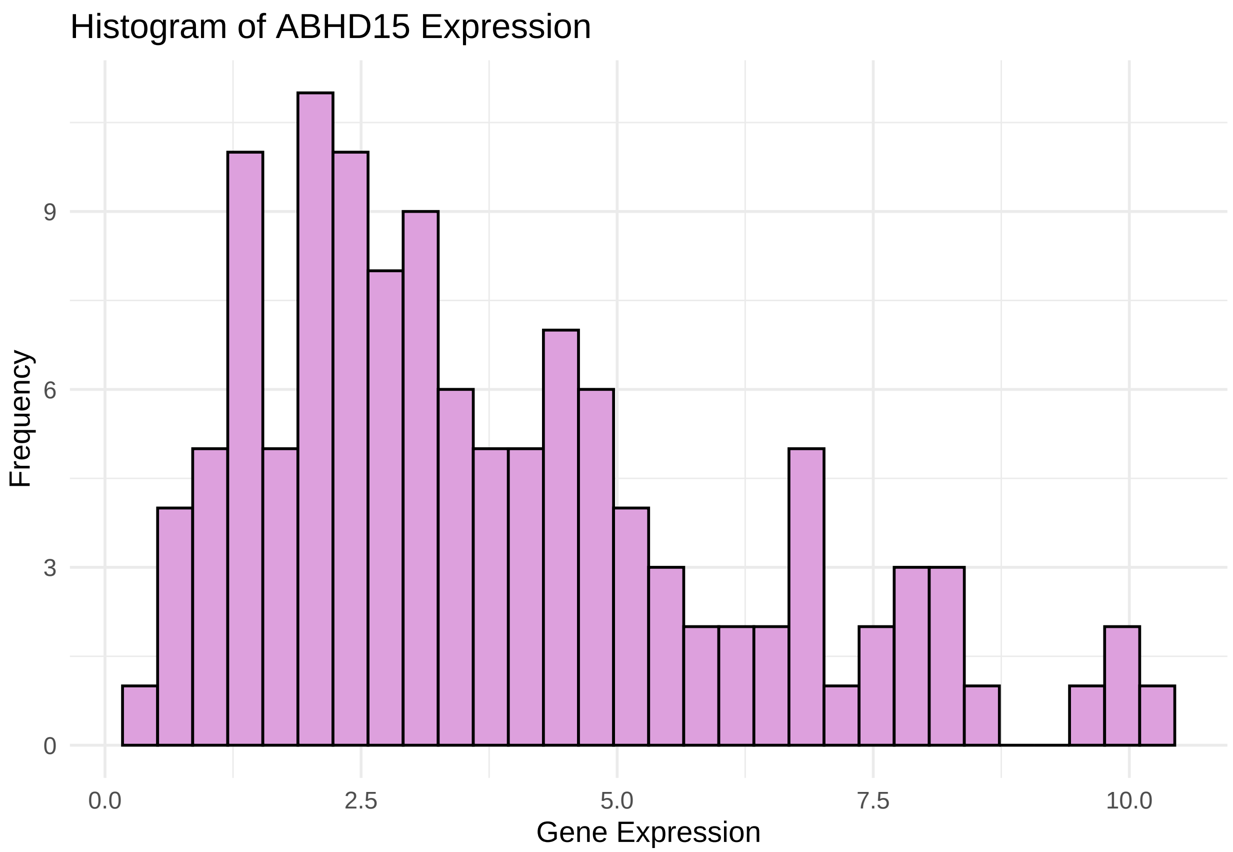


```
my_function <- function(combined_data, gene_list) {

  for (gene_name in gene_list) {

    # Filter for gene of interest
    gene_data <- combined_data %>%
      filter(Gene == gene_name) %>%
      select(participant_id, gene_expression, age, sex, icu_status) %>%
      # Ensure continuous covariate age is numeric
      mutate(age = as.numeric(age)) %>%
      # Cleaning categorical variables
      mutate(across(c(sex, icu_status), ~ str_trim(tolower(.)))) %>%
      #filtering to standardize sex values
      filter(!sex %in% c("unknown", "", "na", "n/a") & !is.na(sex))

    # Histogram
    histogram <- ggplot(gene_data, aes(x = gene_expression)) +
      geom_histogram(fill = "plum", color = "black") +
      labs(title = paste("Histogram of", gene_name, "Expression"),
           x = "Gene Expression", y = "Frequency") +
      theme_minimal()

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

    # Boxplot
    boxplot <- ggplot(gene_data, aes(x = sex, y = gene_expression, fill = icu_status)) +
      geom_boxplot() +
      scale_fill_manual(values=c("hotpink3","lightskyblue")) +
      labs(title = paste("Expression of", gene_name, "by sex"),
           x = "Sex", y = "Gene Expression", fill = 'ICU_status') +
      theme_classic()

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

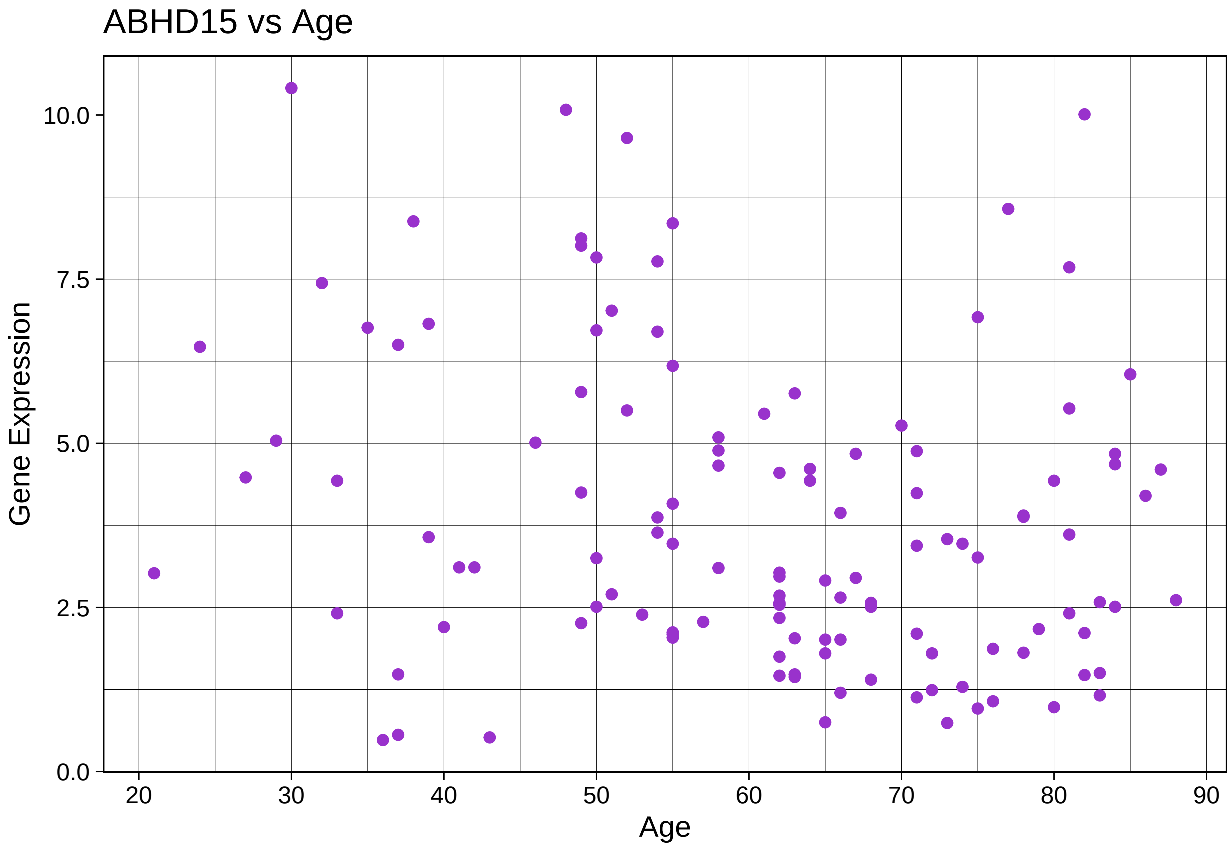
```
my_function(combined_data, gene_list = c("ABHD15","ABI1","AATF"))
```

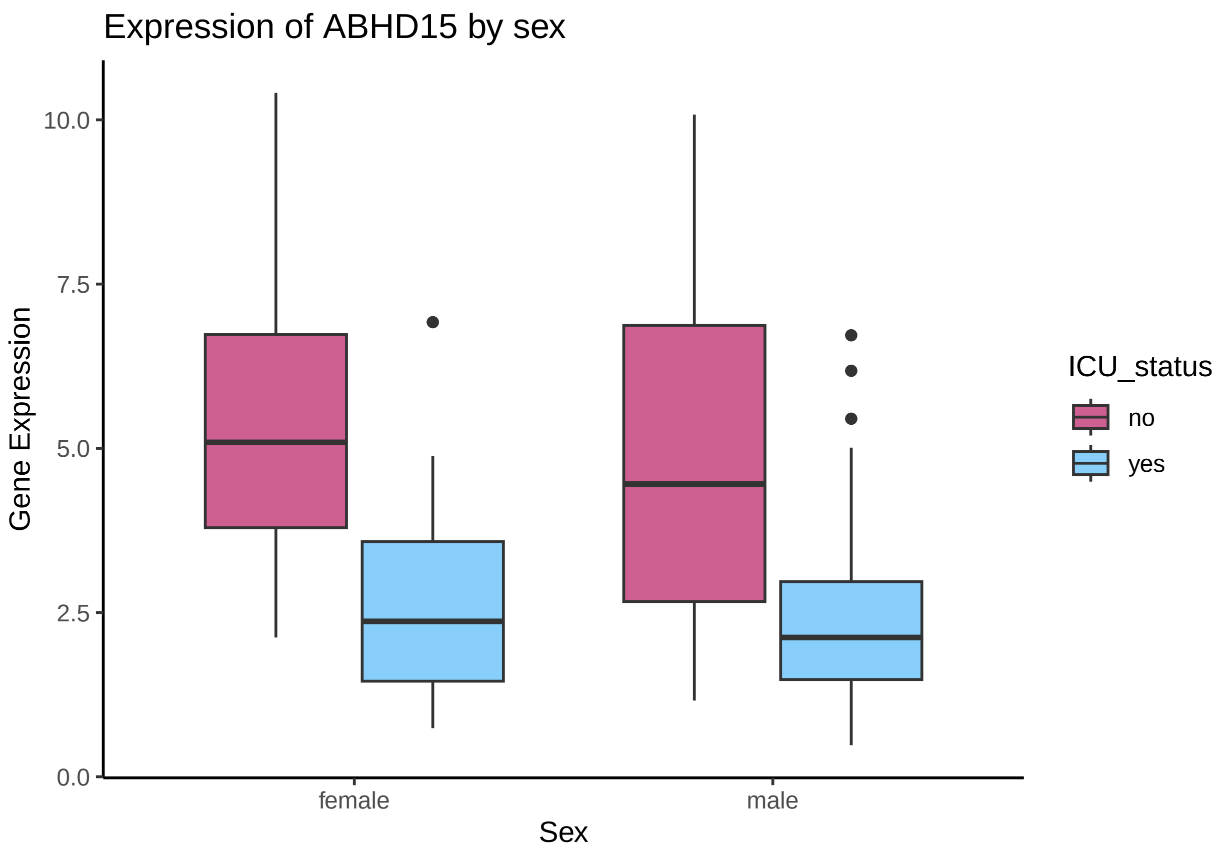
```
## Warning: There was 1 warning in ‘mutate()‘.
## i In argument: ‘age = as.numeric(age)‘.
## Caused by warning:
## ! NAs introduced by coercion
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



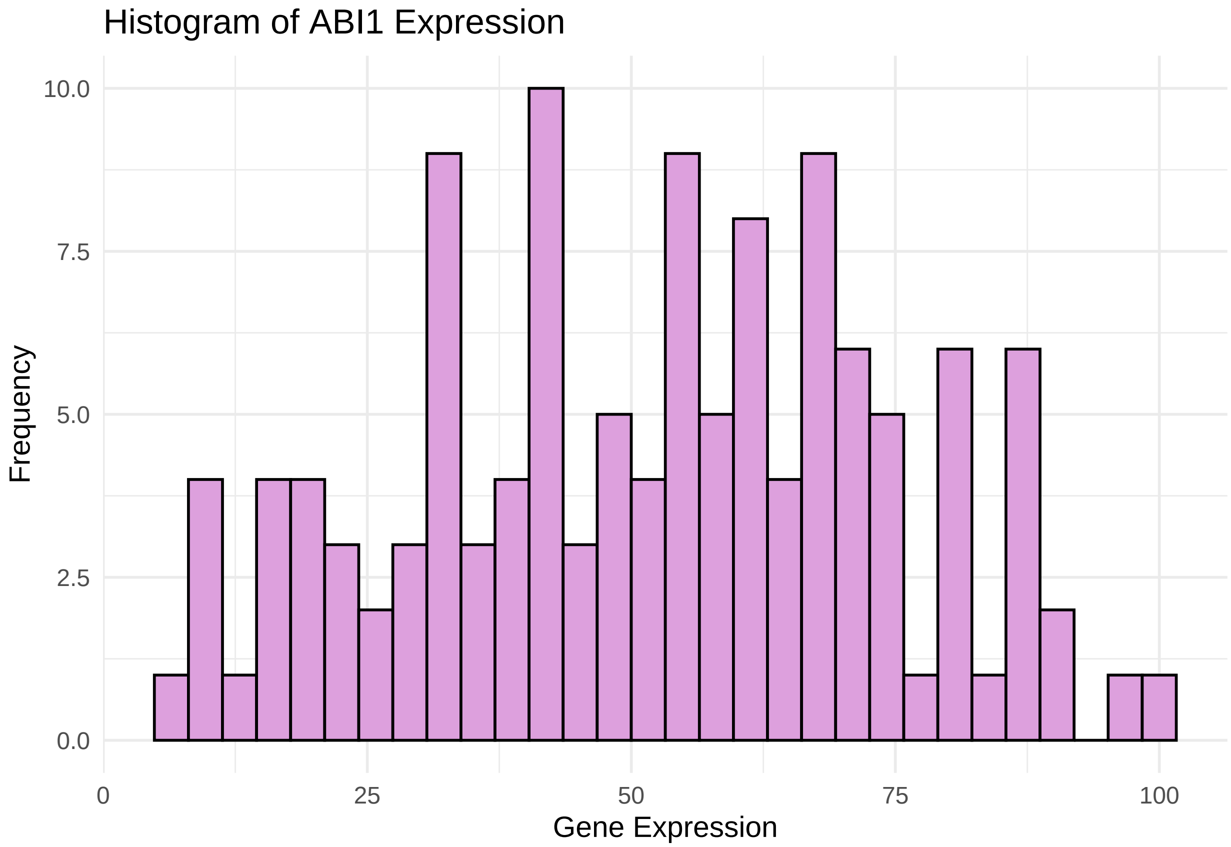
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```


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

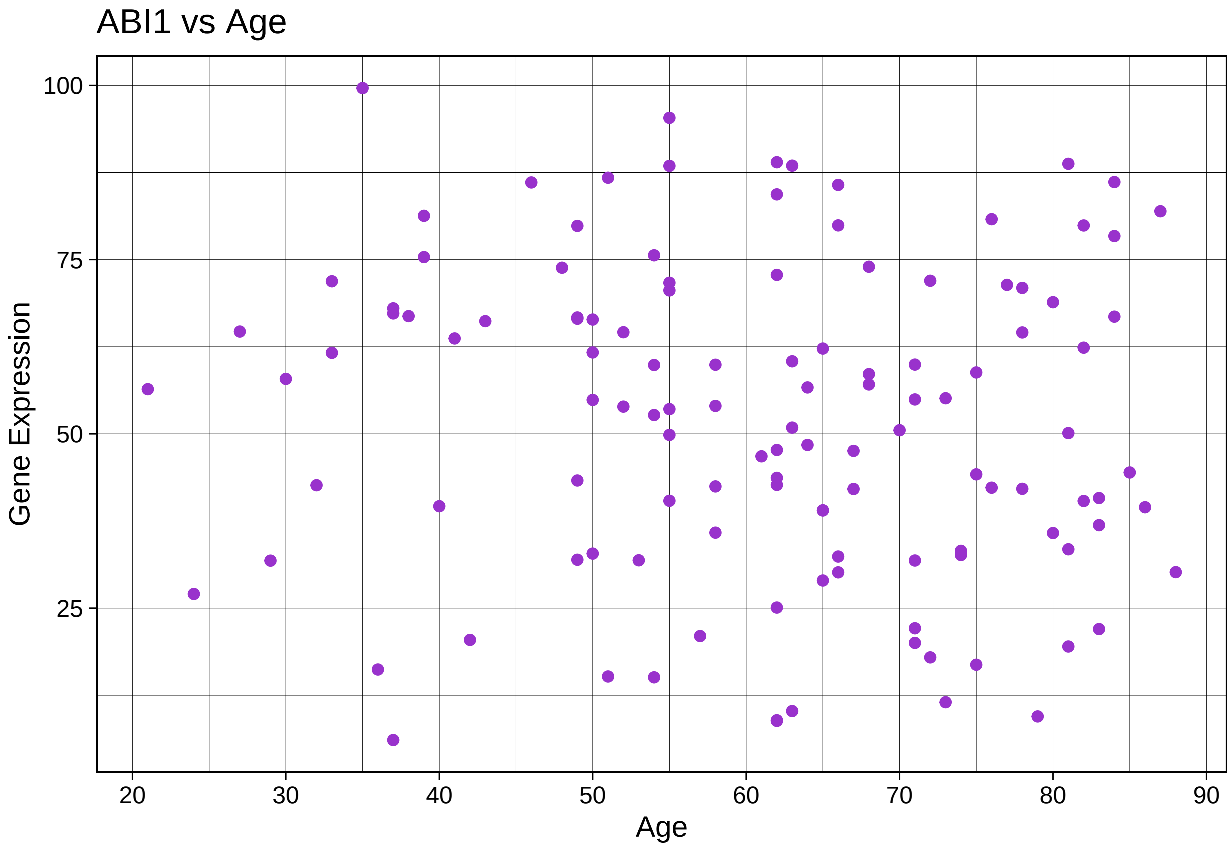


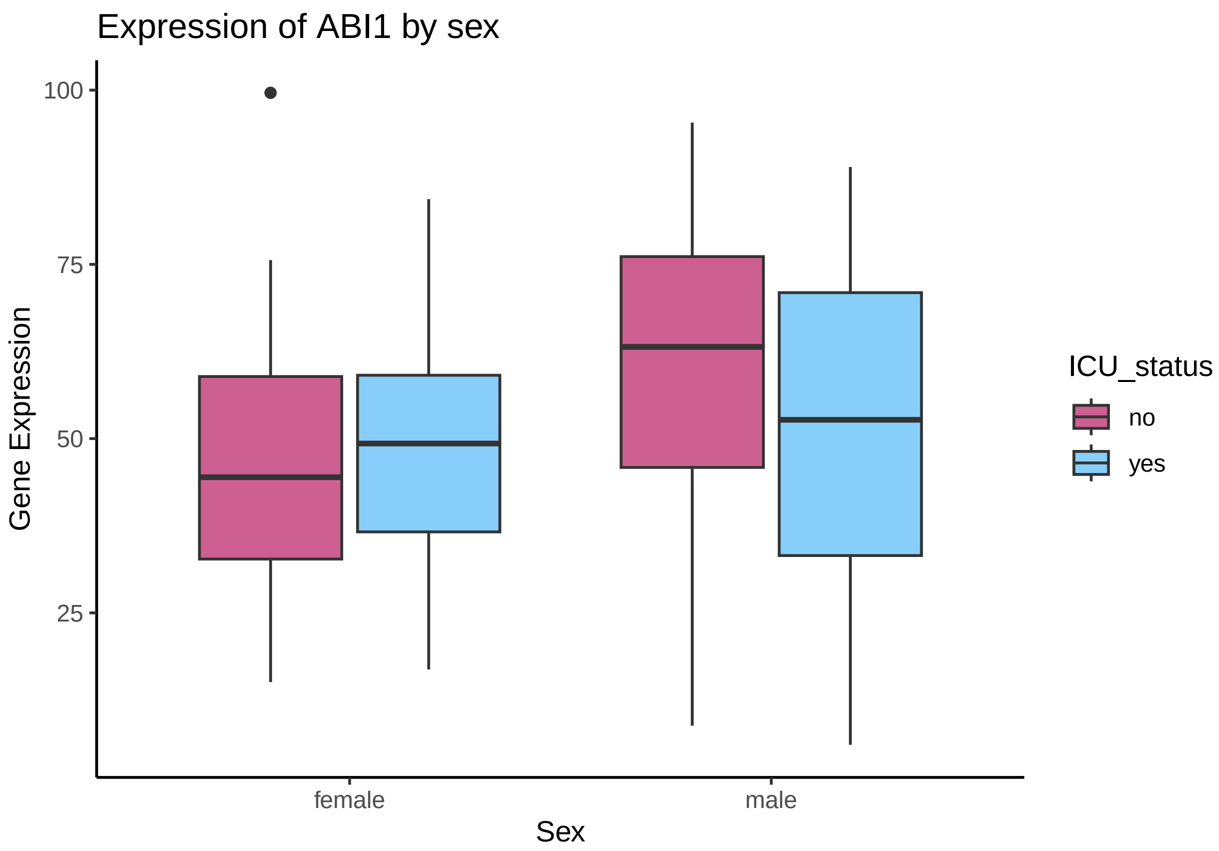
```
## Warning: There was 1 warning in ‘mutate()‘.
## i In argument: ‘age = as.numeric(age)‘.
## Caused by warning:
## ! NAs introduced by coercion
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



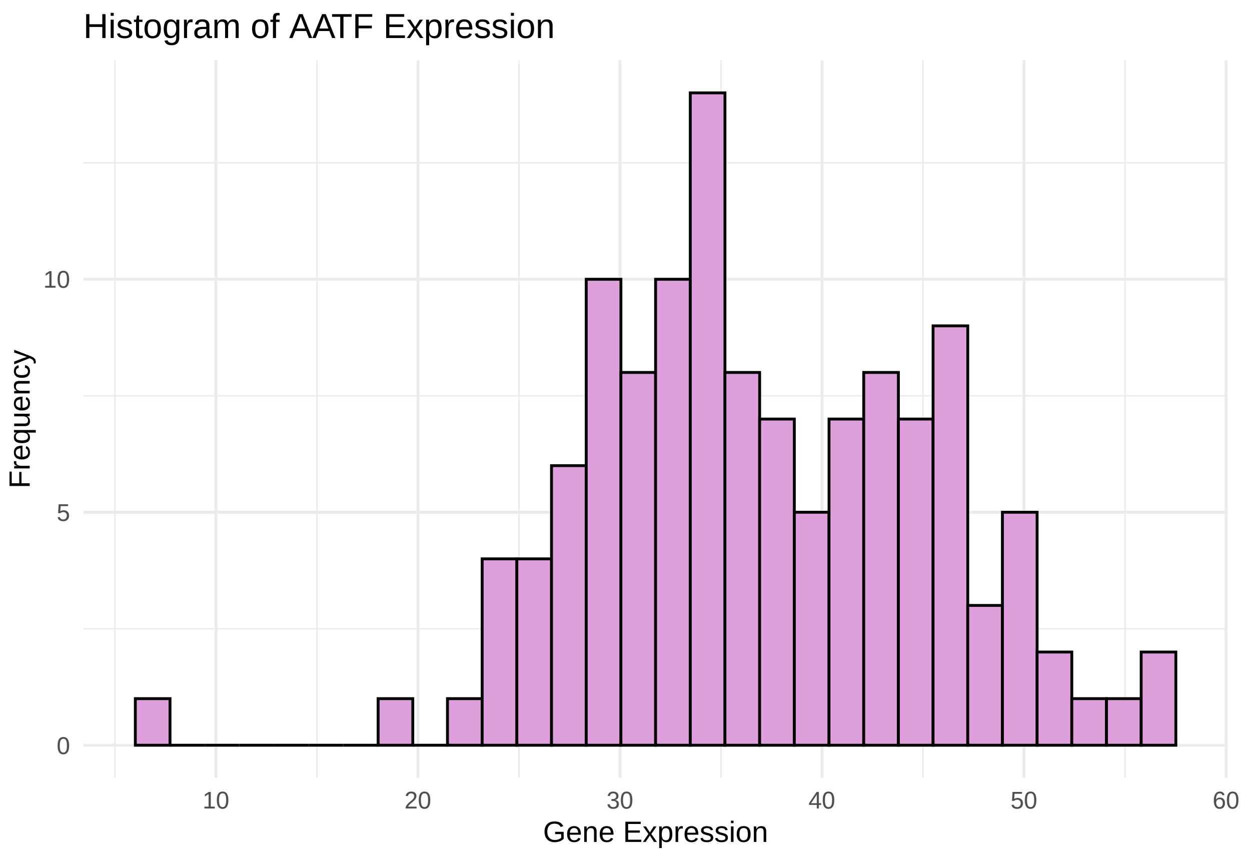
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```


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

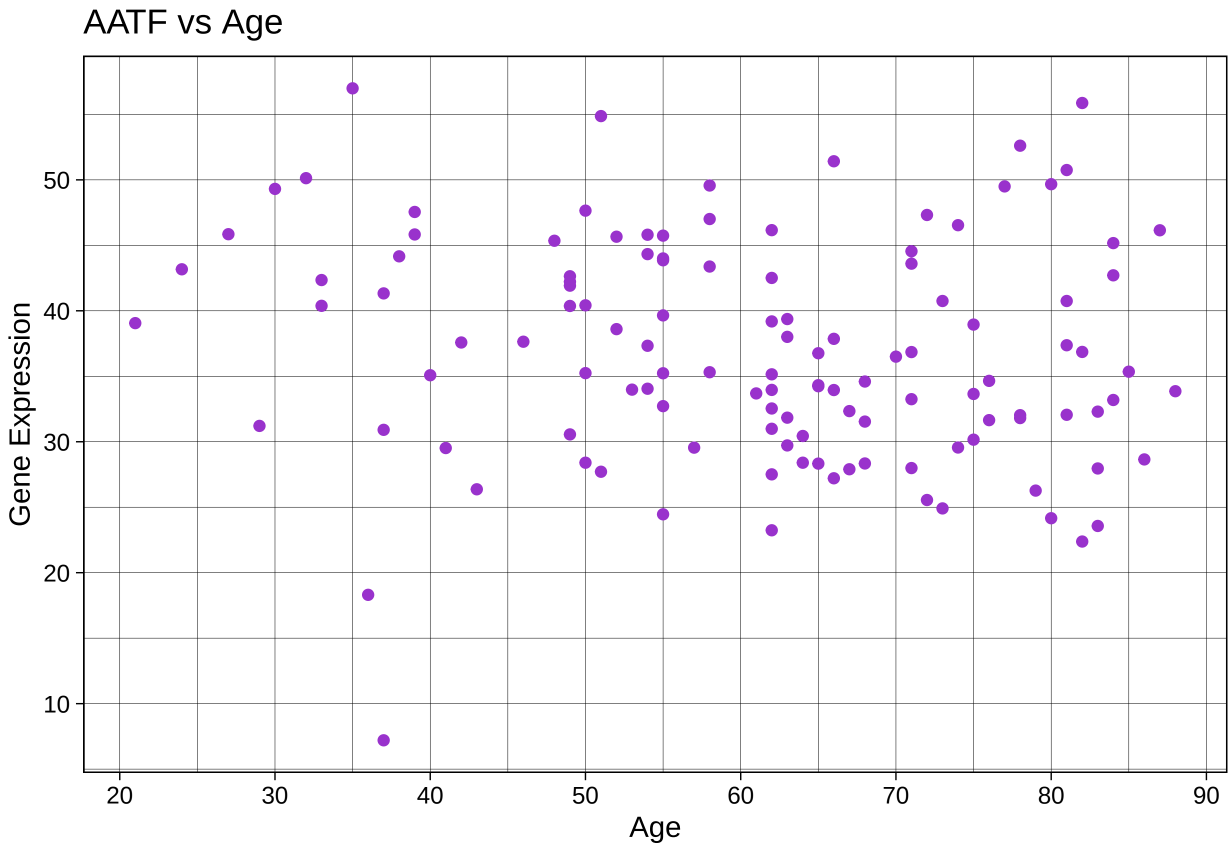


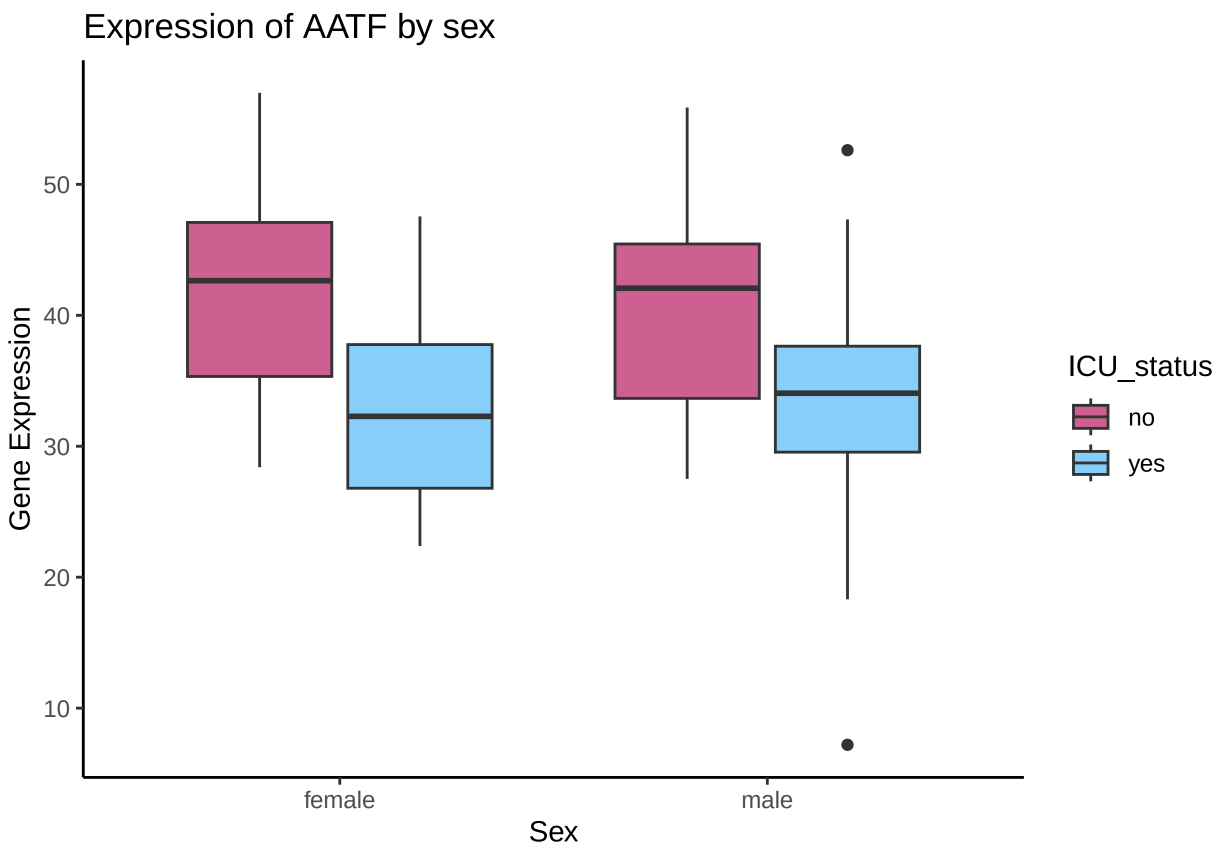
```
## Warning: There was 1 warning in ‘mutate()‘.
## i In argument: ‘age = as.numeric(age)‘.
## Caused by warning:
## ! NAs introduced by coercion
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



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


AATF vs Age
Gene Expression
10
20
30
40
50
Age
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