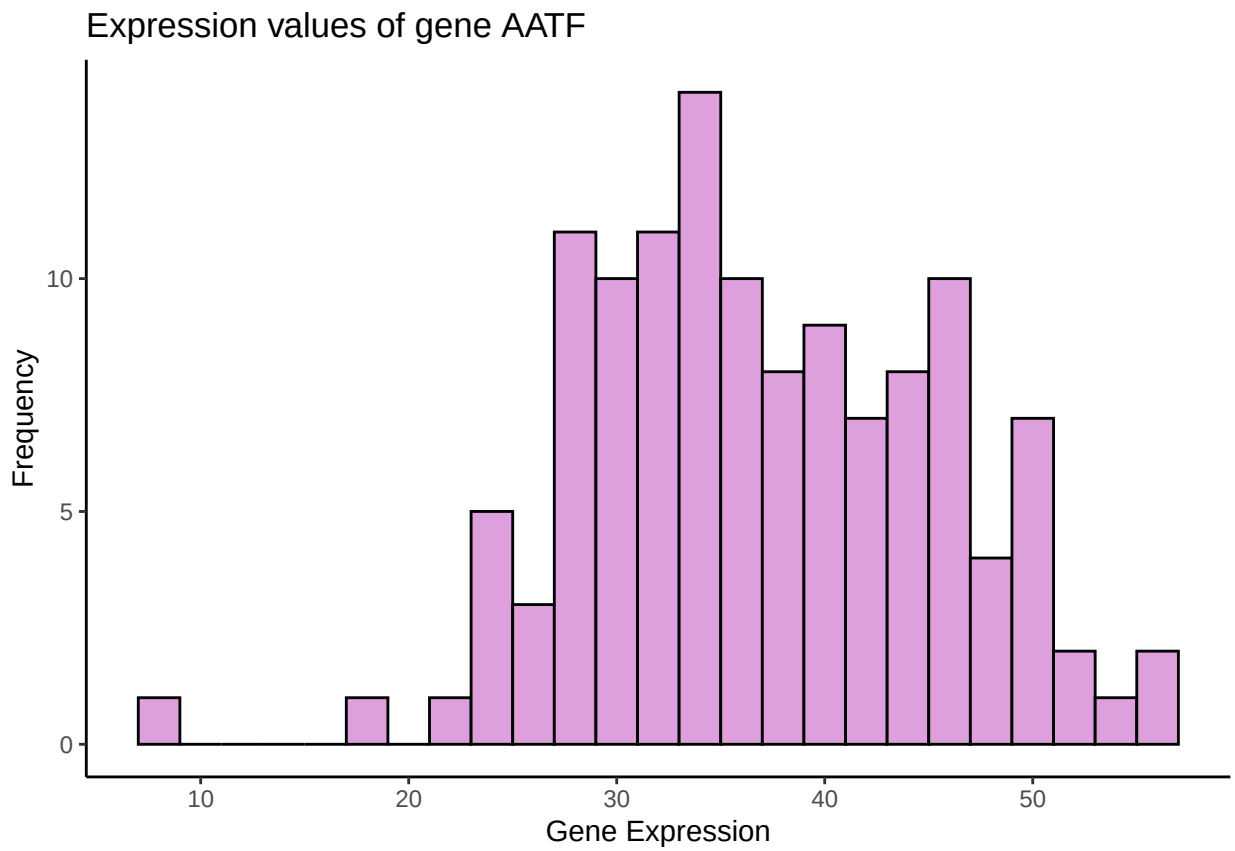
Final project 1

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()
```



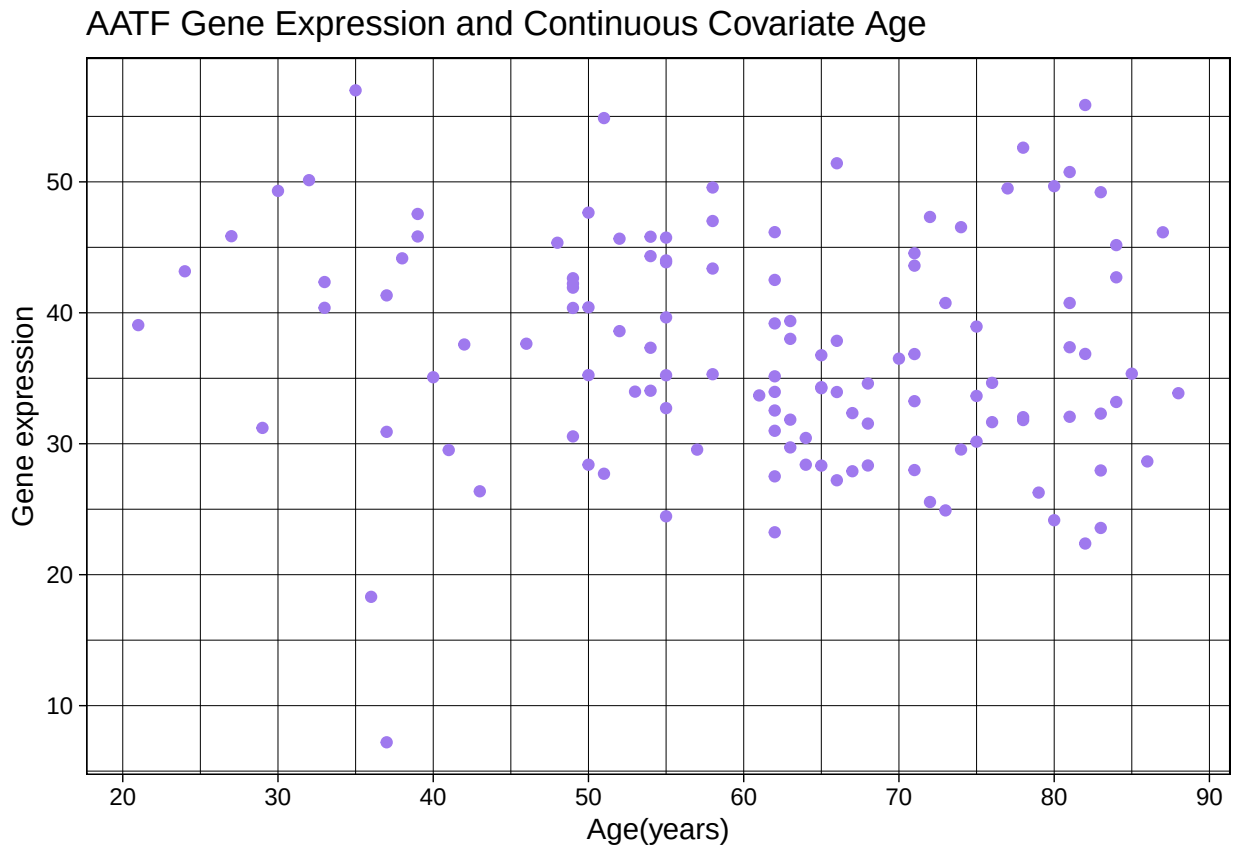
scatter plot

```
#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
```

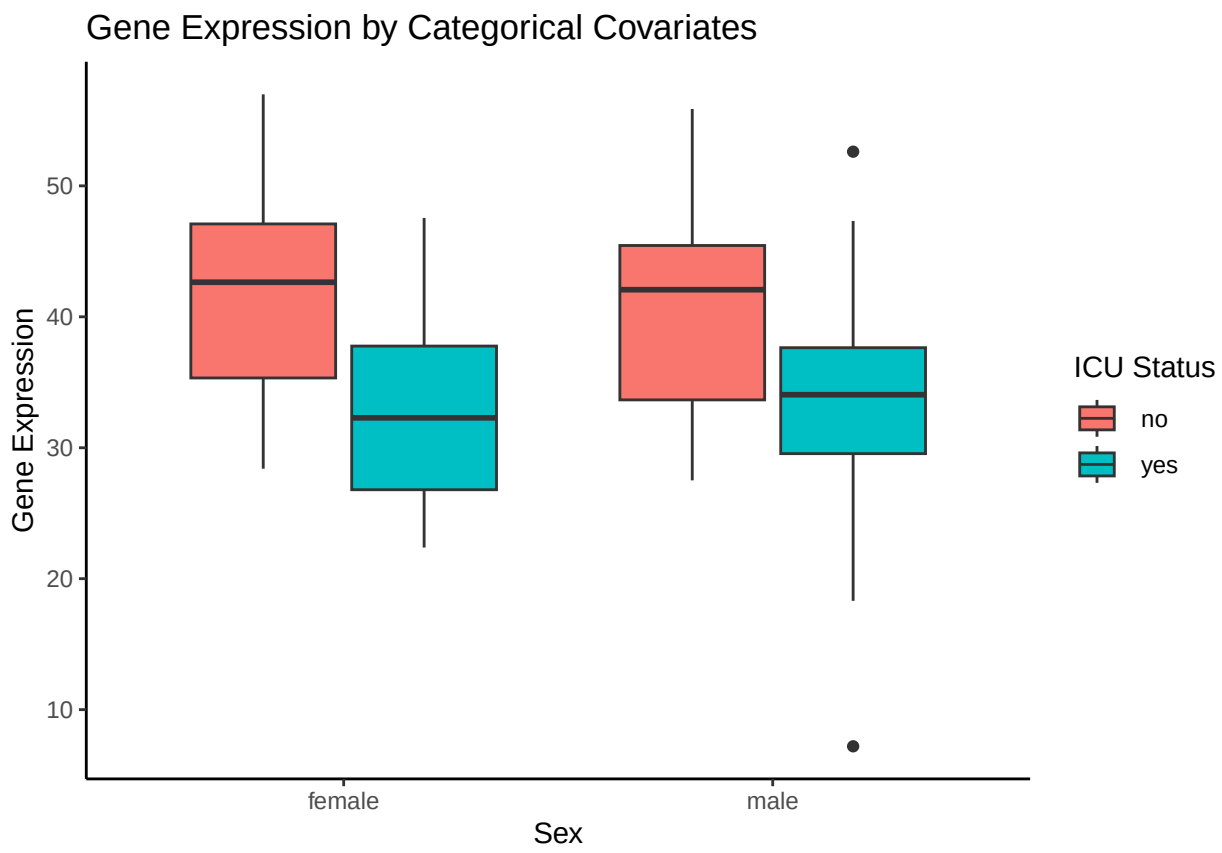
```
#and to ensure my interval breaks take effect properly. used chatgpt and https://vrcacademy.com/tutoria

#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")
```

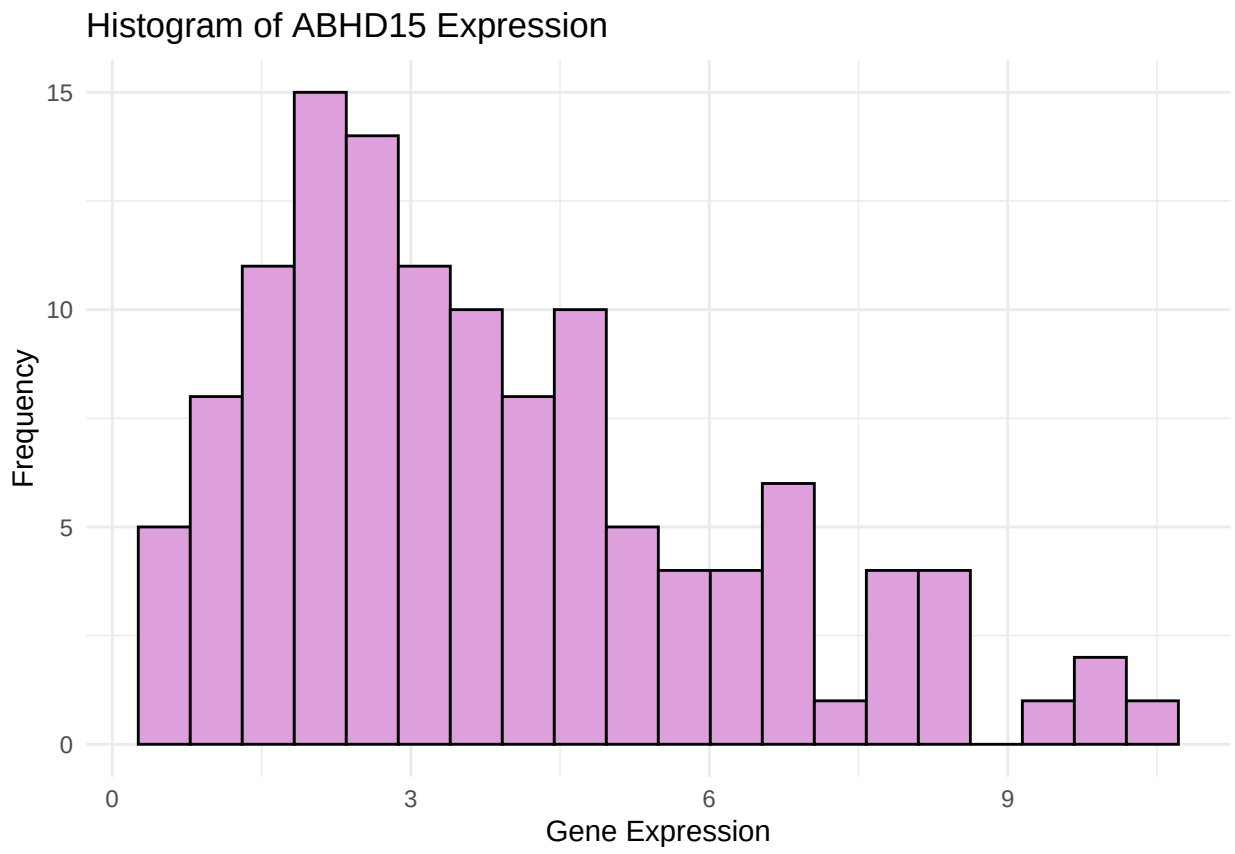



```
my_function <- function(data, gene_list, continuous_cov, categorical_cov1, categorical_cov2,
                        cont_label = continuous_cov,
                        cat1_label = categorical_cov1,
                        cat2_label = categorical_cov2) {

  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
```

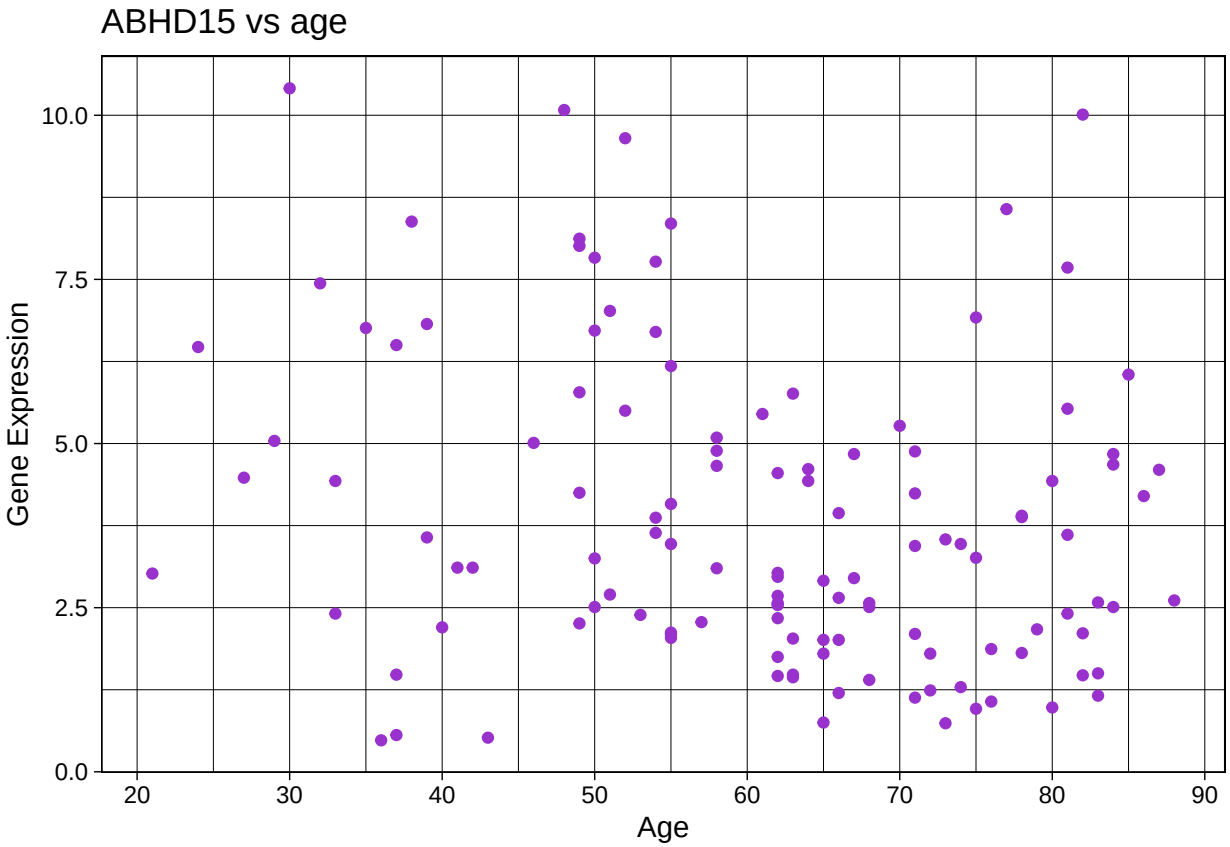
```
    # Histogram
    histogram <- ggplot(gene_data, aes(x = gene_expression)) +
      geom_histogram(fill = "plum", color = "black", bins=20) +
      labs(title = paste("Histogram of", gene_name, "Expression"),
           x = "Gene Expression", y = "Frequency") +
      theme_minimal()

    # 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 = cont_label, 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"),
                        labels = function(x) str_to_title(x)) +
      labs(title = paste("Expression of", gene_name, "by", categorical_cov1),
           x = cat1_label, y = "Gene Expression", fill = cat2_label) +
      scale_x_discrete(labels = function(x) str_to_title(x)) +  # capitalizes x-axis labels
      theme_classic()

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

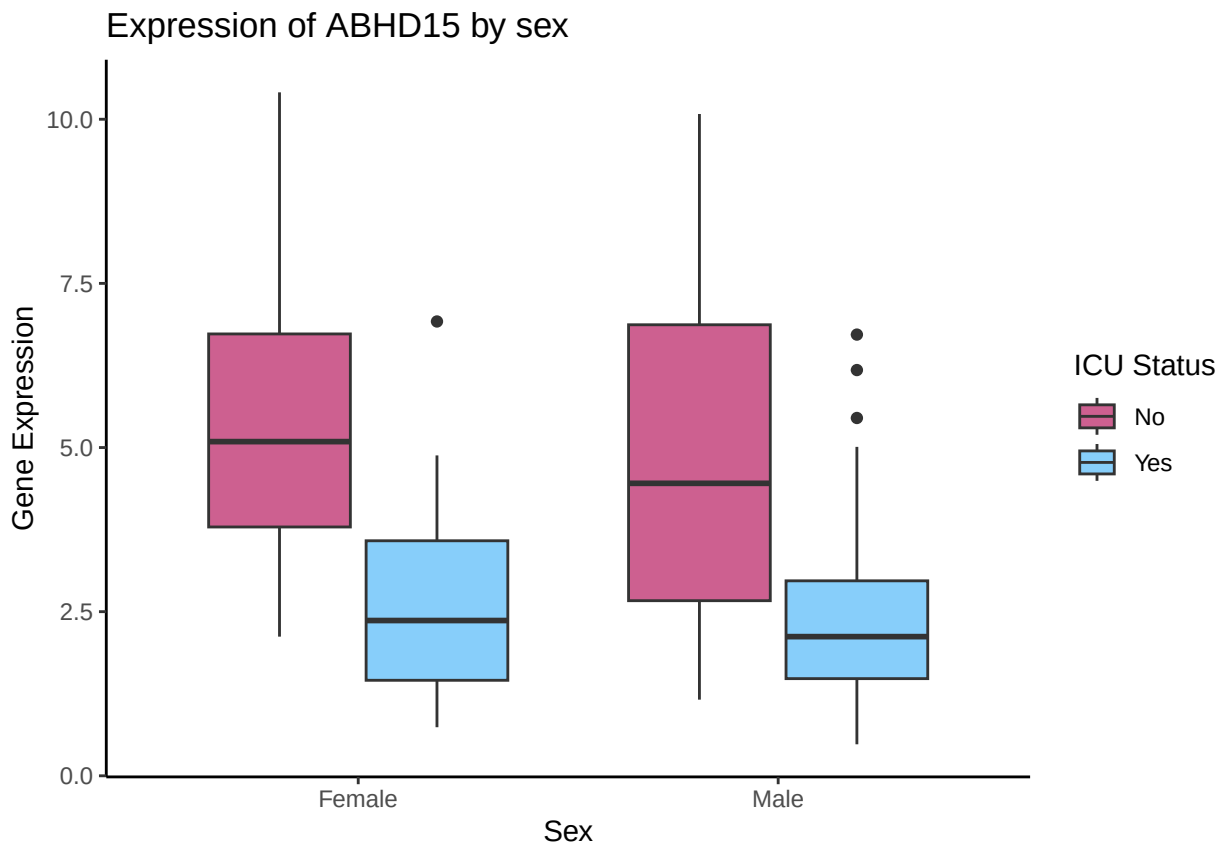
```
#calling my previously defined function and setting values for the parameters
my_function(data = combined_data,
            gene_list = c("ABHD15","ABI1","AATF"),
            continuous_cov = "age",
            categorical_cov1 = "sex",
            categorical_cov2 = "icu_status",
            cont_label = "Age",
            cat1_label = "Sex",
            cat2_label = "ICU Status")
```

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

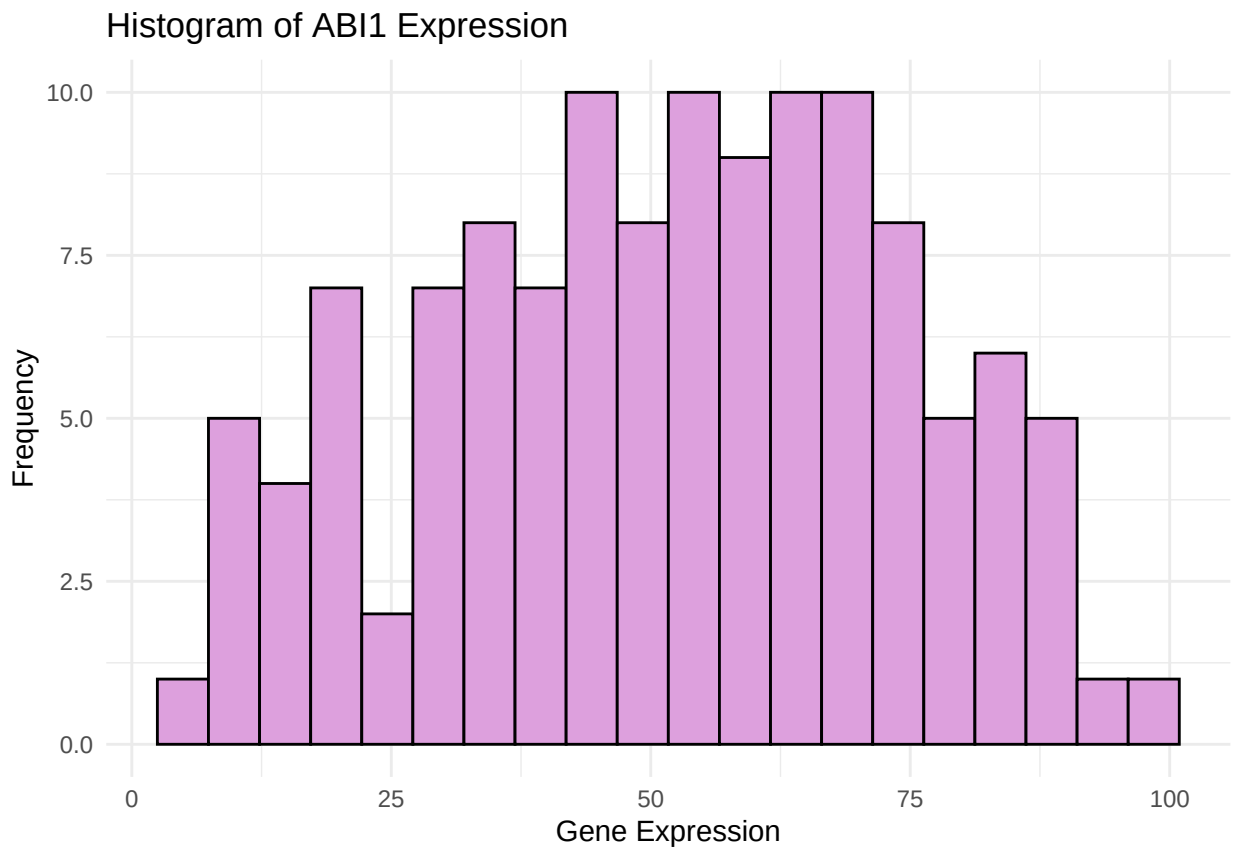



```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

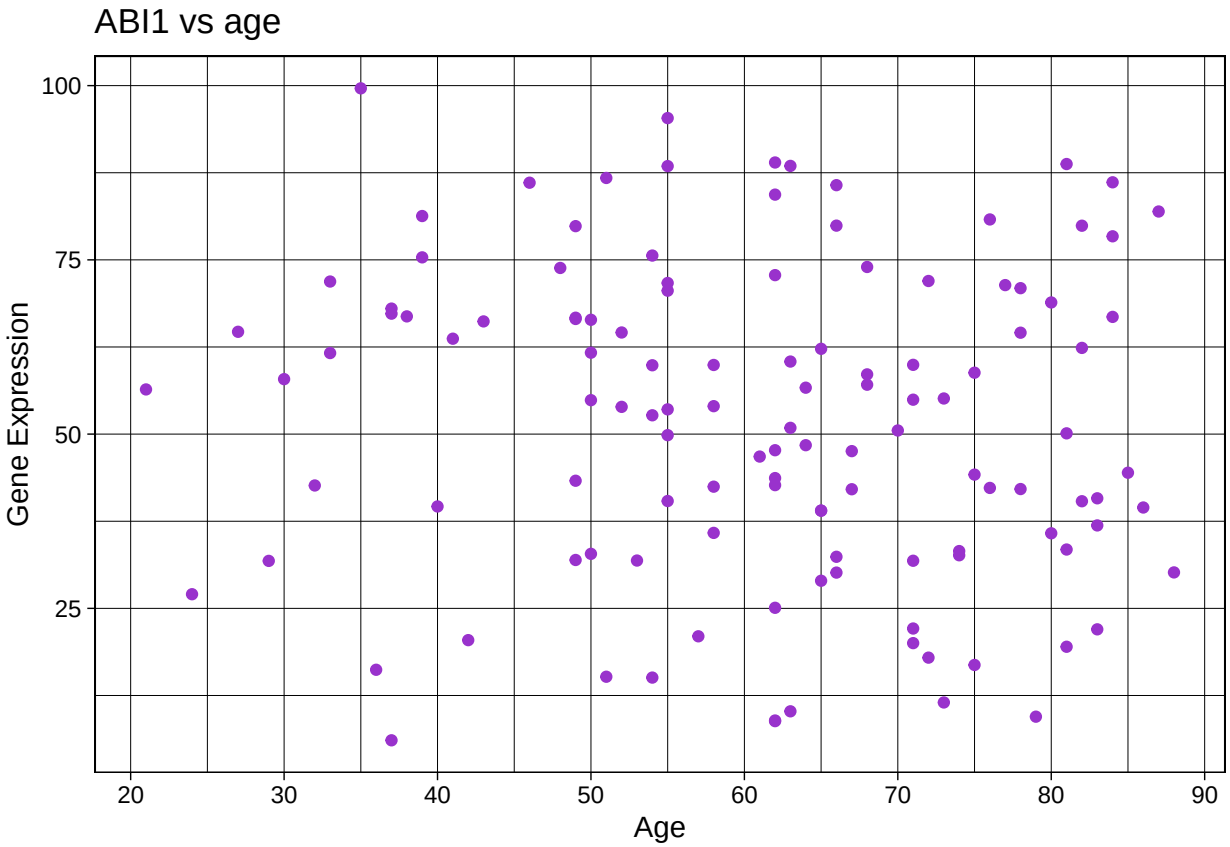

ABHD15 vs age
Gene Expression
Age
0.0
2.5
5.0
7.5
10.0
20
30
40
50
60
70
80
90

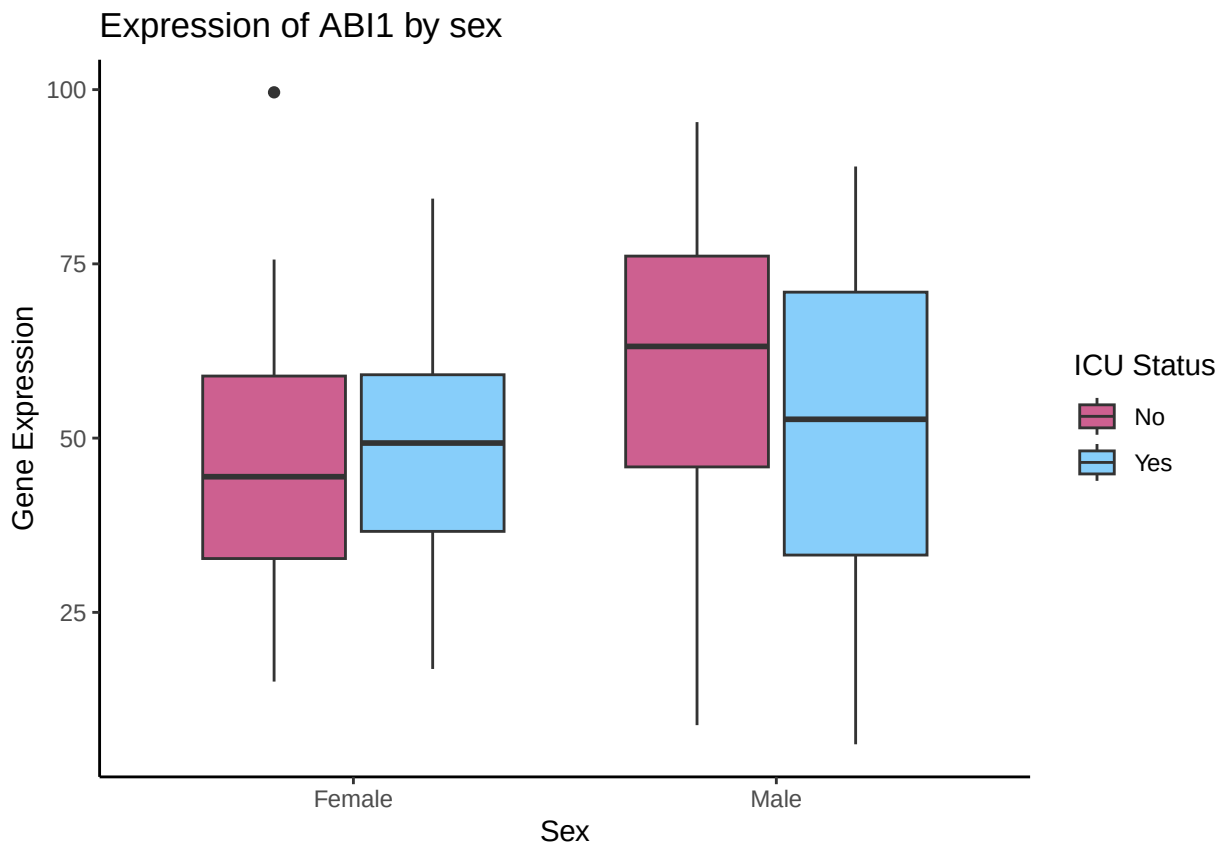



```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `age = as.numeric(age)`.
## Caused by warning:
## ! NAs introduced by coercion
```

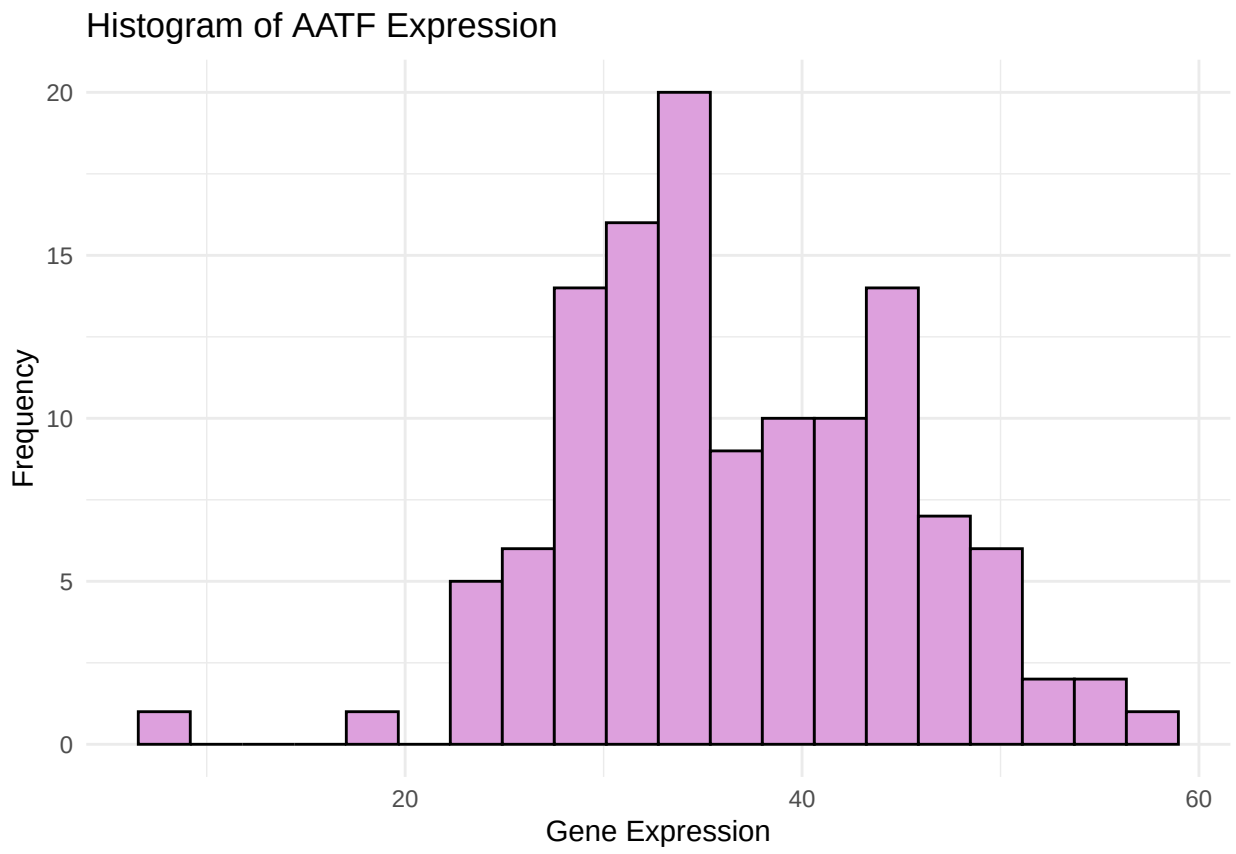



```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

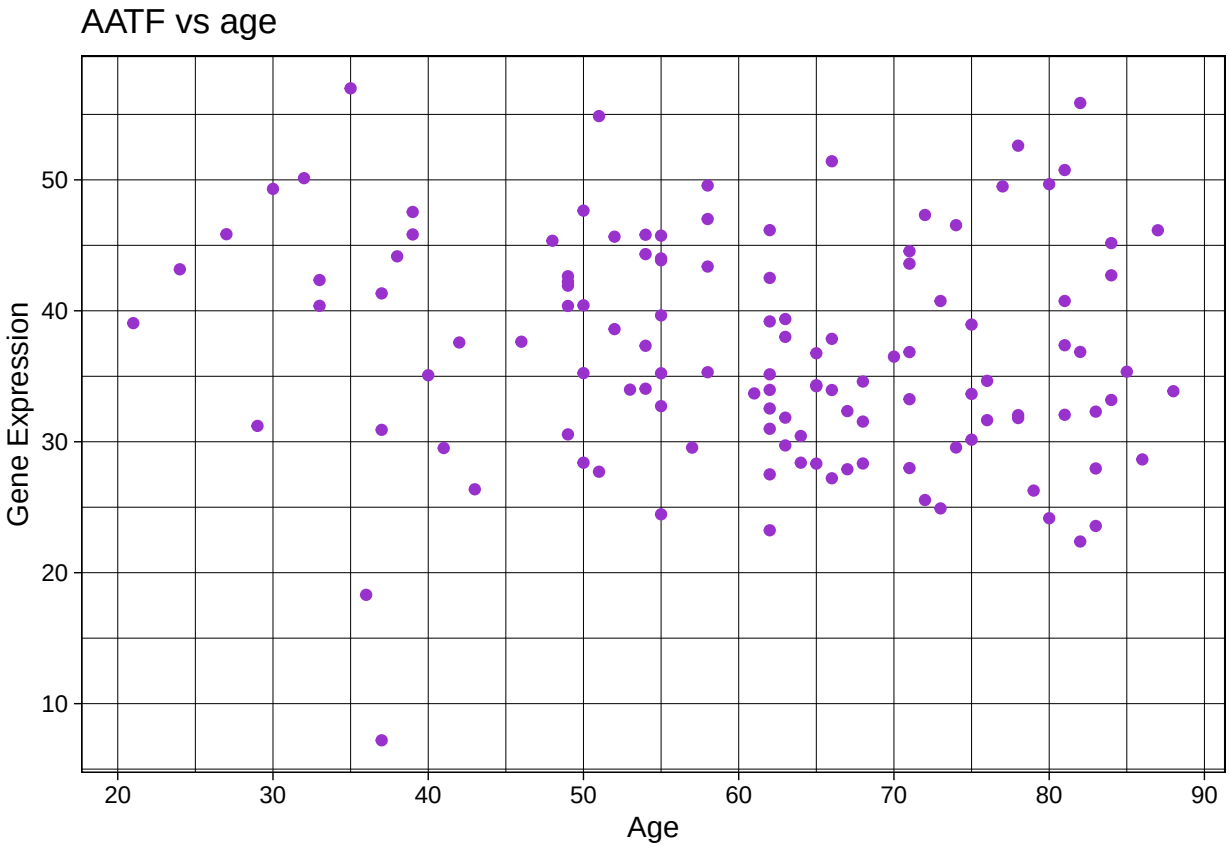

ABI1 vs age
Gene Expression
100
75
50
25
20
30
40
50
60
70
80
90
Age

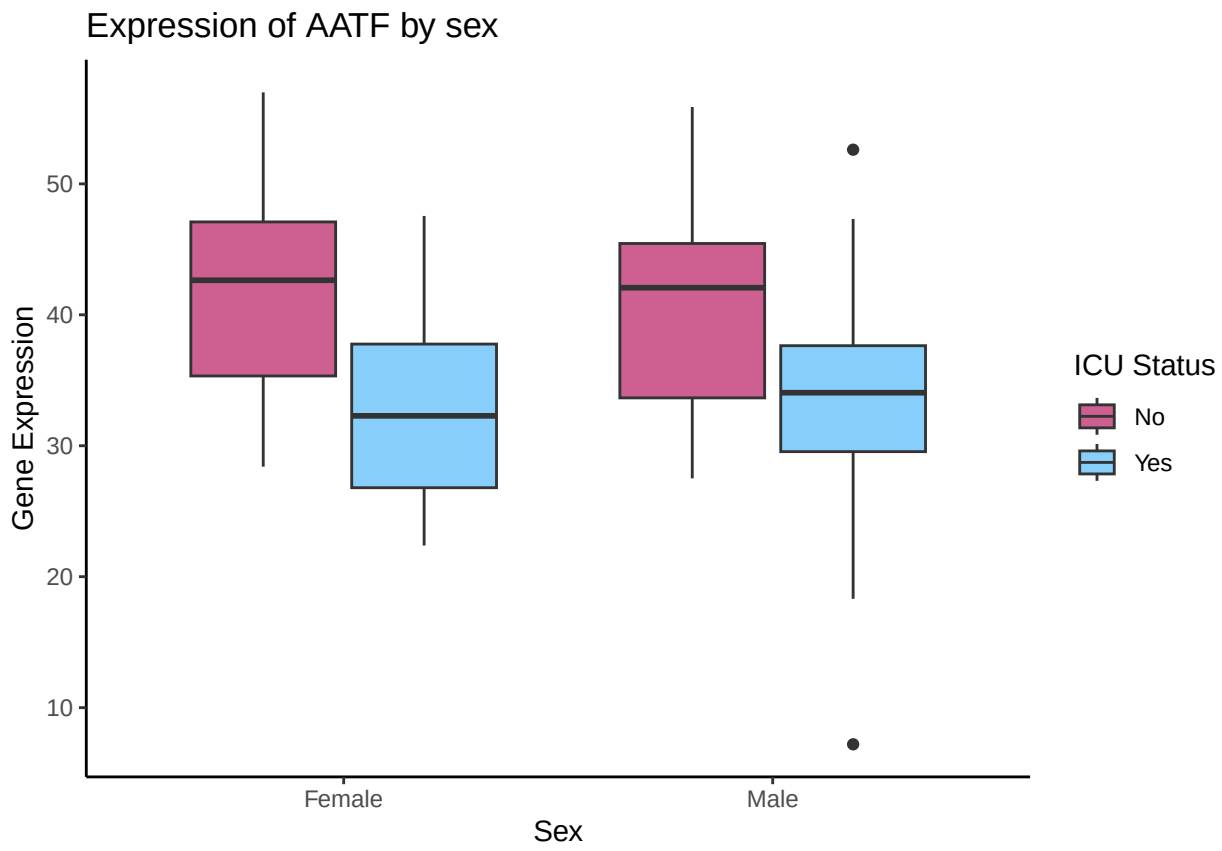



```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `age = as.numeric(age)`.
## Caused by warning:
## ! NAs introduced by coercion
```




```
## 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




```
#cleaning up variables
#chatgpt was used as an aide to create the summary stats table
combined_dataB <- combined_data %>%
  mutate(
    age = as.numeric(age),
    crp = as.numeric(na_if(crp.mg.l., "unknown")),#ensures values are numeric and changes unkowns to na
    ferritin = as.numeric(na_if(ferritin.ng.ml., "unknown")),
    sex = tolower(trimws(sex)),
    icu_status = tolower(trimws(icu_status)),
    mechanical_ventilation = tolower(trimws(mechanical_ventilation))
  ) %>%
  drop_na(crp, ferritin, sex) #gets rid of my unknowns
```

```
## Warning: There were 3 warnings in `mutate()`.
## The first warning was:
## i In argument: `age = as.numeric(age)`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 2 remaining warnings.
```

```
#summary function for my continous variables to calculate mean and sd
cont_summ <- function(x) {
  paste0(round(mean(x, na.rm = TRUE), 1), " (", round(sd(x, na.rm = TRUE), 1), ")")
}
```

```r
#summary function for categorial vars to count n and %
cat_summ <- function(x) {
  tab <- table(x, useNA = "no")
  prop <- prop.table(tab) * 100
  paste0(names(tab), ": ", tab, " (", round(prop, 1), "%)", collapse = "; ")
}

#creating my summary table stratified by patient sex
summary_table <- combined_dataB %>%
  group_by(sex) %>%
  summarise(
    Age = cont_summ(age),
    CRP = cont_summ(crp),
    Ferritin = cont_summ(ferritin),
    "ICU Status" = cat_summ(icu_status),
    "Mechanical Ventilation" = cat_summ(mechanical_ventilation) ,
    ) %>%
  rename(Sex = sex)

# Format into latex table
kable(summary_table, format = "latex", booktabs = TRUE,
      caption = "Summary Statistics Stratified by Sex") %>%
  kable_classic() %>%
  kable_styling(latex_options = c("hold_position", "scale_down")) #%>%
```

Table 1: Summary Statistics Stratified by Sex

Sex	Age	CRP	Ferritin	ICU Status	Mechanical Ventilation
female	61.5 (17.3)	115.4 (98.3)	632 (1076.3)	no: 2300 (52.3%); yes: 2100 (47.7%)	no: 3000 (68.2%); yes: 1400 (31.8%)
male	62.1 (14.7)	146.7 (102.2)	1014.2 (1021.7)	no: 2300 (38.3%); yes: 3700 (61.7%)	no: 2900 (48.3%); yes: 3100 (51.7%)

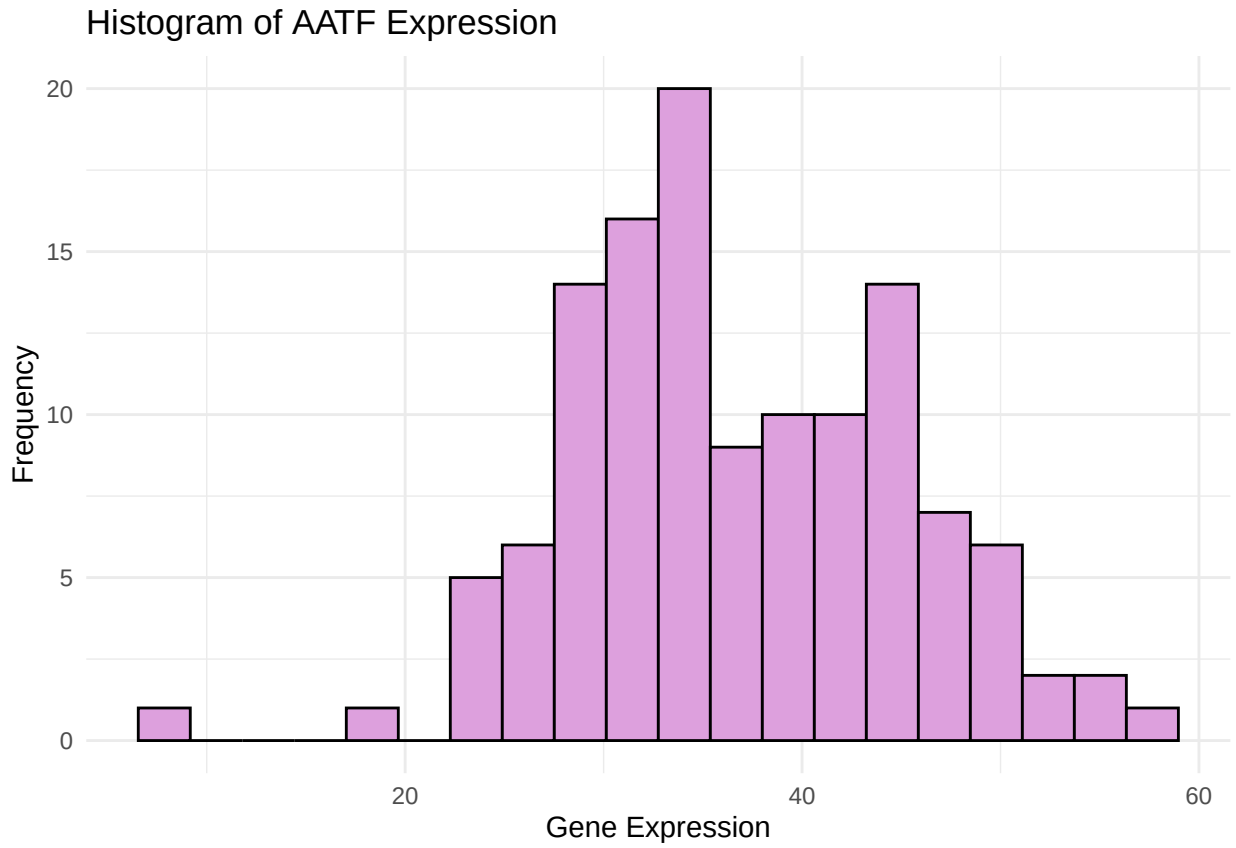
```r
  #save_kable("summary_table.tex")
print(summary_table)
```

```
## # A tibble: 2 x 6
##   Sex    Age         CRP           Ferritin  `ICU Status` Mechanical Ventilati~1
##   <chr>  <chr>       <chr>         <chr>     <chr>        <chr>
## 1 female 61.5 (17.3) 115.4 (98.3)  632 (107~ no: 2300 (5~ no: 3000 (68.2%); yes~
## 2 male   62.1 (14.7) 146.7 (102.2) 1014.2 (~ no: 2300 (3~ no: 2900 (48.3%); yes~
## # i abbreviated name: 1: `Mechanical Ventilation`
```

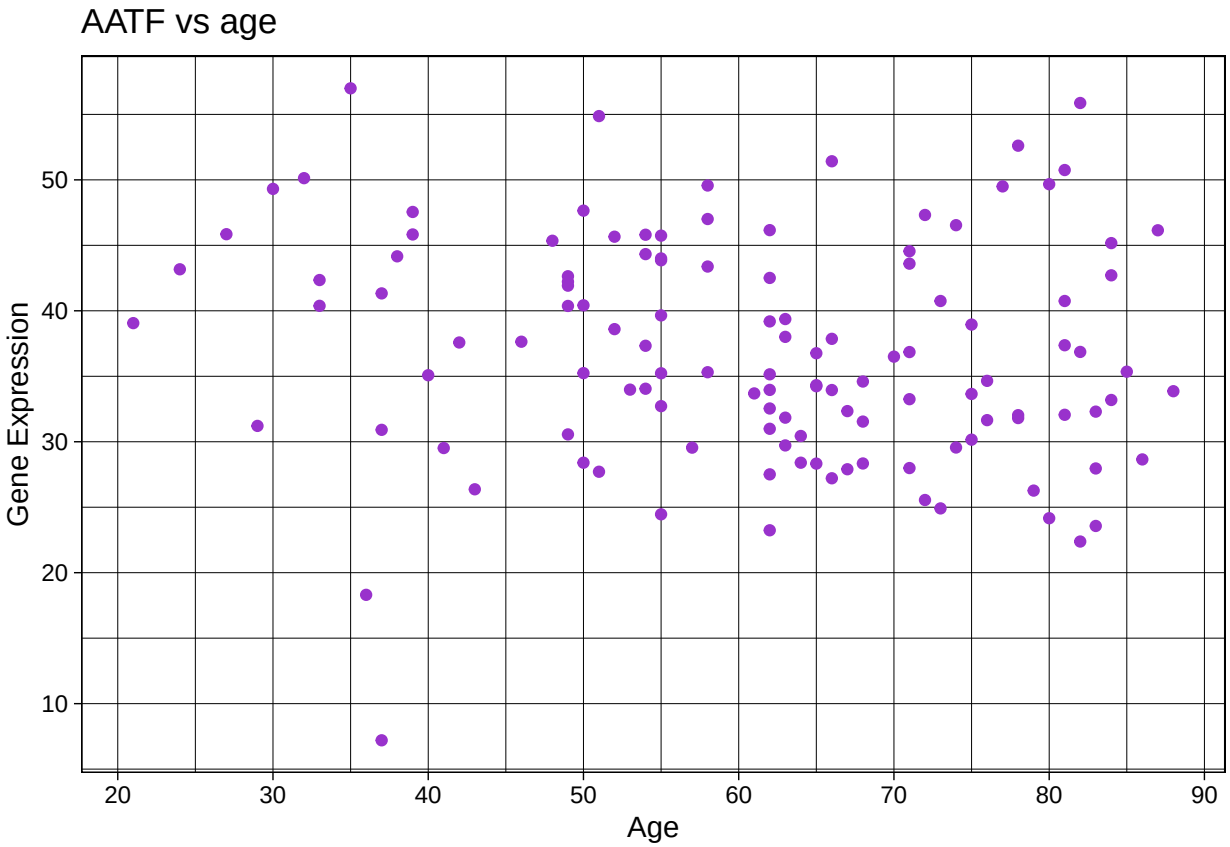
Plots for Gene of Interest

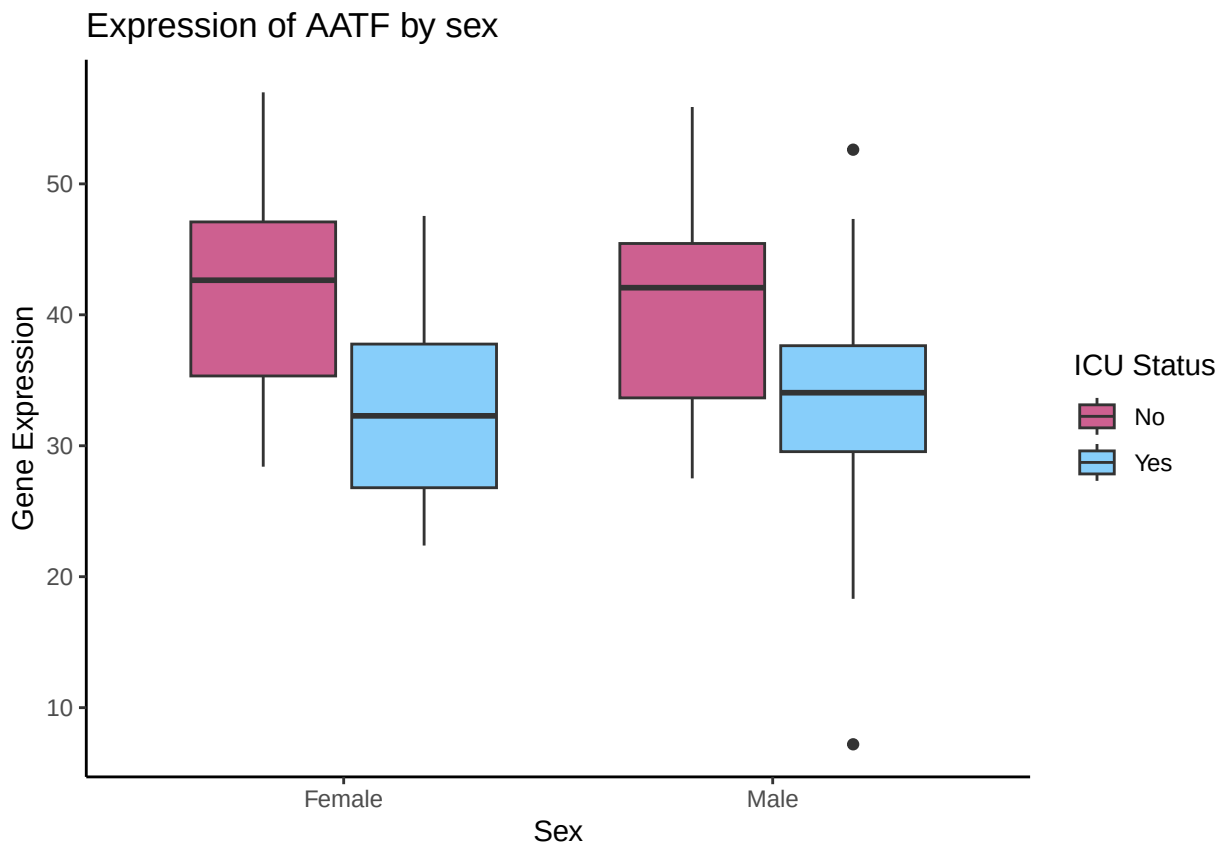
```r
#using previously defined function to generate final publication ready plots
my_function(data = combined_data,
            gene_list = c("AATF"),
            continuous_cov = "age",
            categorical_cov1 = "sex",
            categorical_cov2 = "icu_status",
            cont_label = "Age",
            cat1_label = "Sex",
            cat2_label = "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()‘).
```

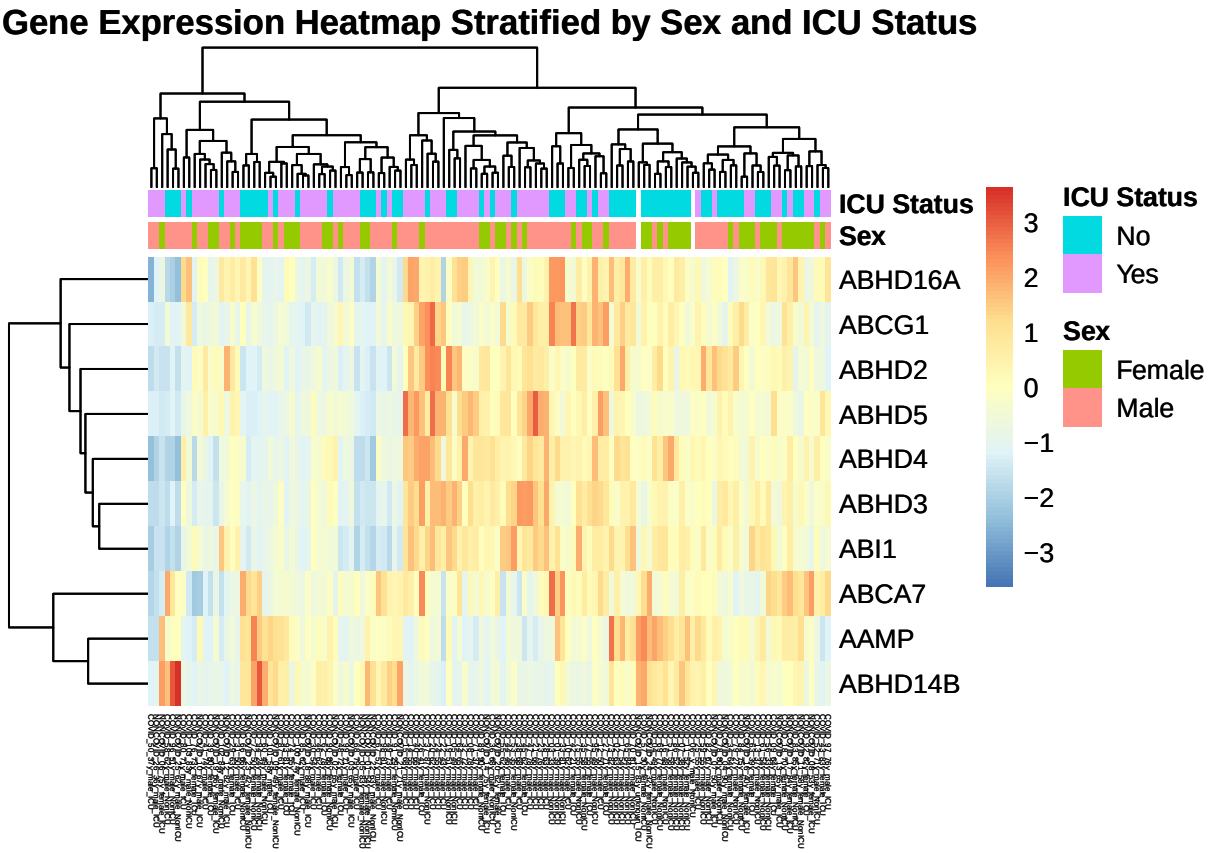

AATF vs age
Gene Expression
Age
10
20
30
40
50
20
30
40
50
60
70
80
90



Heatmap

```
#chatgpt was used to assist in the creation of my heatmap
#Calculate variance of each gene
variance <- apply(genedata[,-1], 1, var)

#Orders by decreasing variance
genedata_ordered <- genedata[order(variance, decreasing = TRUE), ]

# Select top 10 variable genes
top_genes <- genedata_ordered[1:10, ]

# Format data for pheatmap
# Remove "Gene" column so rows are numeric
rownames(top_genes) <- top_genes$Gene
top_genes_matrix <- as.matrix(top_genes[,-1])


# Prepare metadata annotations
rownames(metadata) <- metadata$participant_id
metadata_for_heatmap <- metadata %>%
  filter(tolower(trimws(sex)) != "unknown") %>%
  select("Sex" = sex, "ICU Status" = icu_status) %>%
  mutate(Sex = str_to_title(Sex),
    `ICU Status` = str_to_title(`ICU Status`))
```

```r
#Generate heatmap
heatmap <- pheatmap(top_genes_matrix,
         clustering_distance_rows = "euclidean",
         clustering_distance_cols = "euclidean",
         annotation_col = metadata_for_heatmap,
         main = "Gene Expression Heatmap Stratified by Sex and ICU Status",
         scale = "row",
         fontsize_col = 3)
print(heatmap)
```

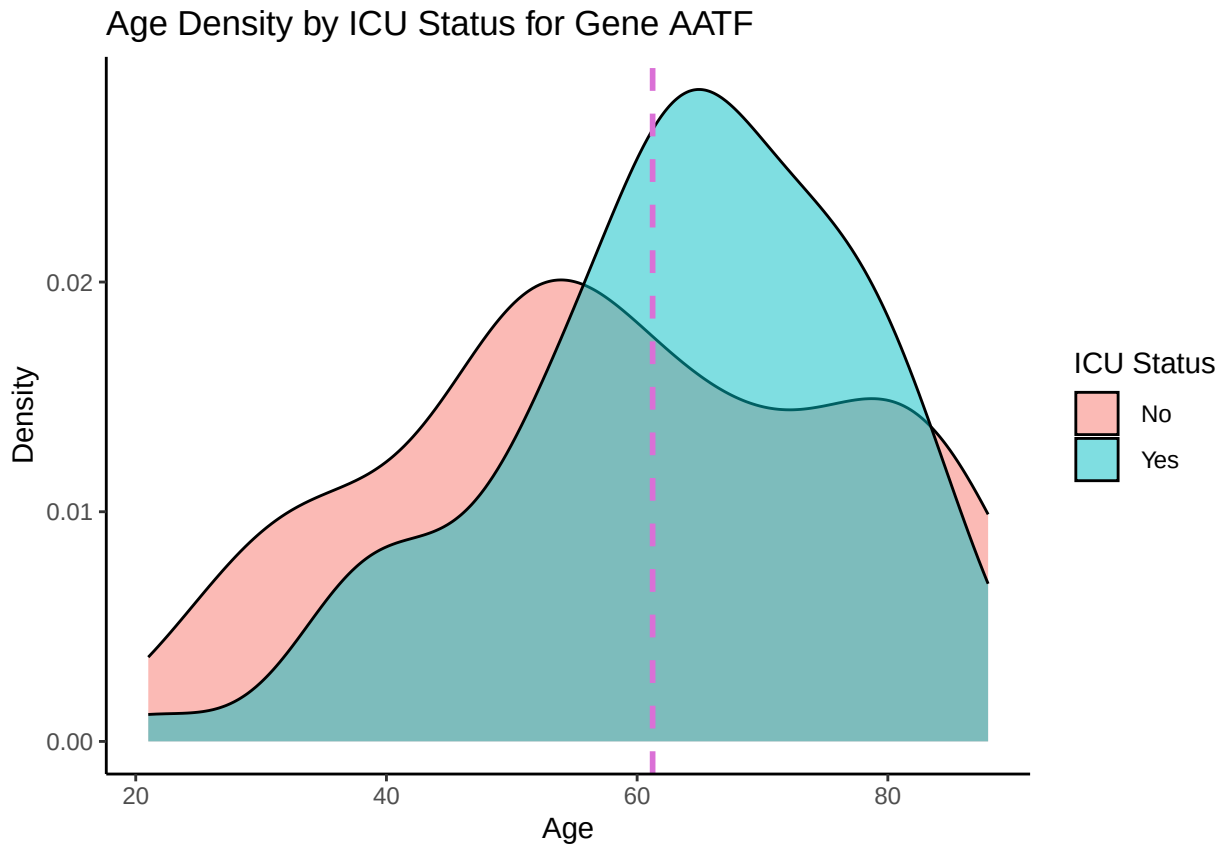


Density Plot

```r
#Resources used when selecting new plot type to use
#https://www.sthda.com/english/wiki/ggplot2-density-plot-quick-start-guide-r-software-and-data-visualiz
#chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://web.stanford.edu/class/stats32/assets/lect
densityplot <- ggplot(AATFData %>%
           mutate(icu_status_clean = str_to_title((icu_status))),
         aes(x = age, fill = icu_status_clean)) +
          geom_density(alpha = 0.5) +        # semi-transparent to see data overlaps
          labs(title = "Age Density by ICU Status for Gene AATF",
         x = "Age", y = "Density",
         fill= 'ICU Status') +
  geom_vline(xintercept = mean(AATFData$age, na.rm = TRUE),
    color = "orchid", linetype = "dashed", linewidth = 1
  ) +
```

```
  theme_classic()
densityplot
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (‘stat_density()‘).
```




```
#ggsave("densityplot.png", plot = densityplot, width=6, height=4)
#ggsave("heatmap.png", plot =heatmap, width=8, height=6)
#citation("dplyr")
#citation("readr")
#citation("tidyverse")
#citation("ggplot2")
#citation("kableExtra")
#citation("pheatmap")
#citation("stringr")
```