103_Project_Submission2

Rob Ross-Shannon

2024-08-06

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.1

##
## 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
```

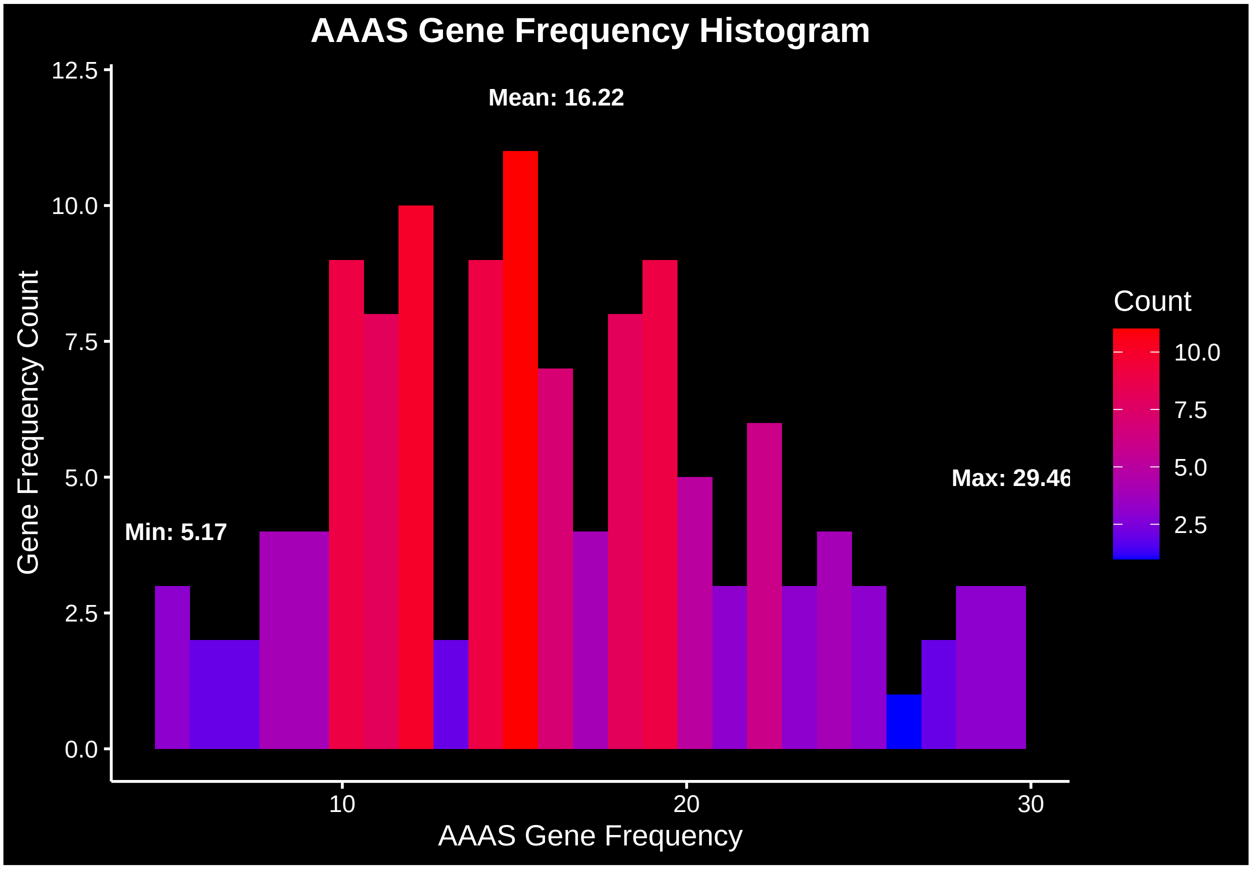
```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.1

## Warning: package 'ggplot2' was built under R version 4.4.1

## Warning: package 'tidyr' was built under R version 4.4.1

## Warning: package 'readr' was built under R version 4.4.1

## Warning: package 'forcats' was built under R version 4.4.1

## Warning: package 'lubridate' was built under R version 4.4.1

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

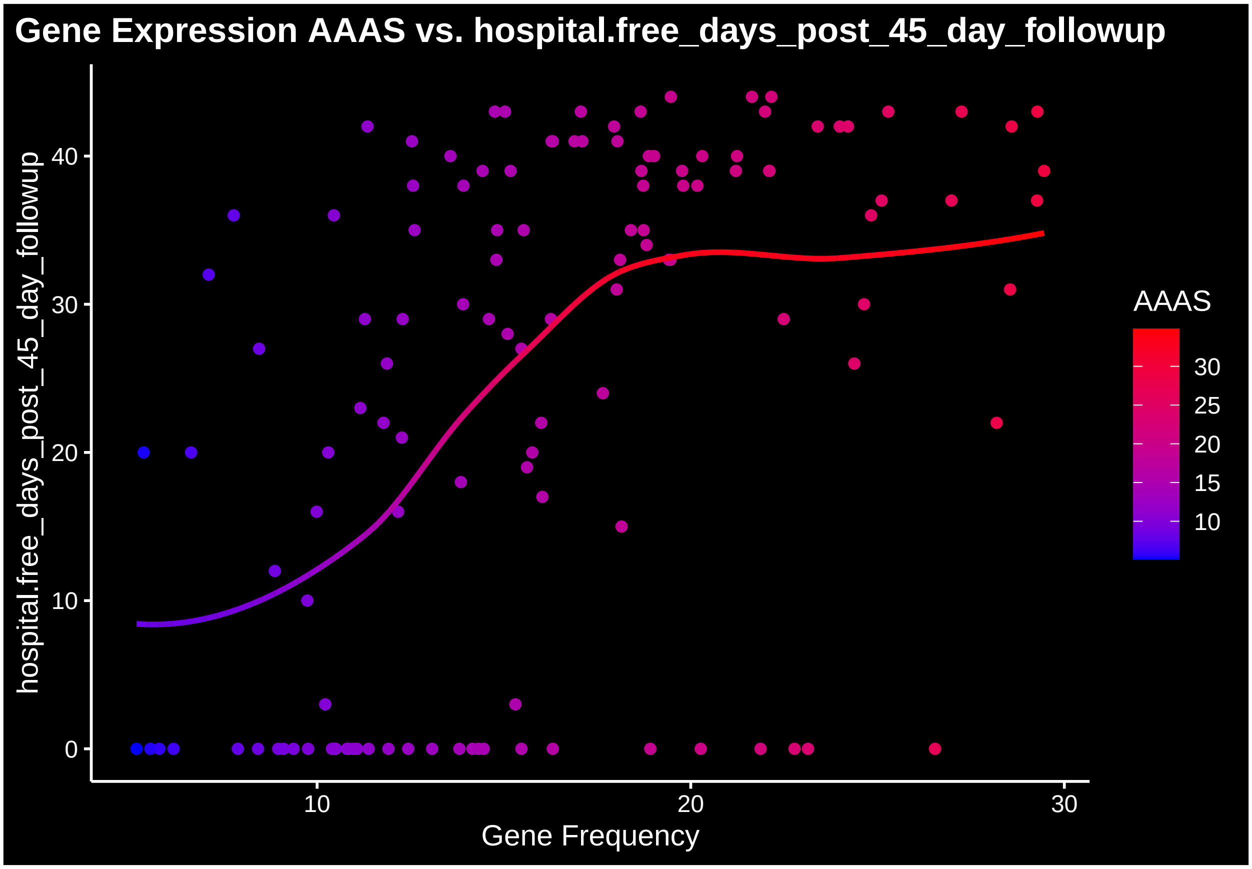
```
#Reading in data sets
setwd("/Users/Rob Ross-Shannon/Documents/GitHub/103_Project")
genes <- read.csv("~/Documents/GitHub/103_Project/QBS103_GSE157103_genes.csv")
series_matrix <- read.csv("~/Documents/GitHub/103_Project/QBS103_GSE157103_series_matrix.csv")

#Cleaning data set (row names)
rownames(genes) <- genes$X
genes <- select(genes, -c('X'))

#Transposing data set and renaming index
genesT <- as.data.frame(t(genes))
genesT$participant_id <- rownames(genesT)

#Combing the data sets into one data frame
combinedData <- inner_join(genesT, series_matrix, by = "participant_id")
```

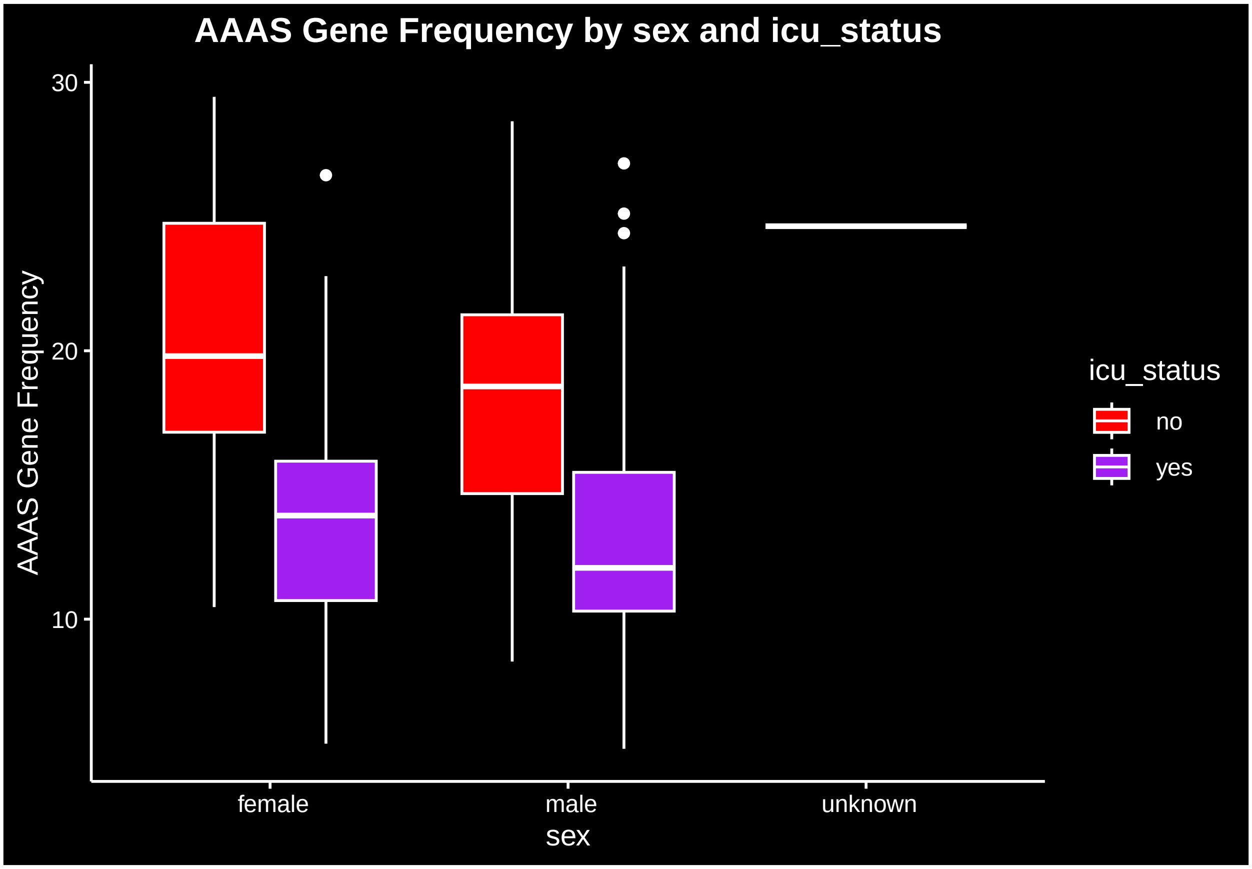
```
project_plots <- function(df_name, gene_name, cont_covariate, cat_covariate1, cat_covariate2){

  #Filtering just for inputs of interest
  cleanedData <- select(df_name, c(gene_name, cont_covariate, cat_covariate1, cat_covariate2))
   #Creating project theme for plots
  project_theme <- theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5, face = "bold"),
    # Define my axis
    title = element_text(colour = "white"),
    axis.line = element_line(colour = "white", linewidth = rel(1)),
    axis.title = element_text(colour = "white"),
    axis.text = element_text(color = "white"),
    axis.ticks = element_line(colour = "white"),
    # Set plot background
    plot.background = element_rect(fill = "black"),
    panel.background = element_blank(),
    legend.key = element_blank(),
    legend.text = element_text(colour = "white"),
    legend.background = element_rect(fill = "black"),
    legend.title = element_text(colour = "white", ),
    # Move legend
    legend.position = 'right')

  #Code to create histogram
  annotations <- data.frame(
    x = c(round(min(cleanedData[[gene_name]]), 2), round(mean(cleanedData[[gene_name]]), 2),
          round(max(cleanedData[[gene_name]]), 2)),
    y = c(4, 12, 5),
    label = c("Min:", "Mean:", "Max:"))

  #Creating histogram
  histogram <- ggplot(cleanedData,aes_string(x = gene_name)) +
    geom_histogram(aes(fill = ..count..), bins = 25)+
    #Creating gradient color representation
    scale_fill_gradient("Count", low = "blue", high = "red")+
    geom_text(data = annotations, aes(x = x, y = y, label = paste(label, x)), size = 3,
              fontface = "bold", color = "white")+
    labs(x = paste0(gene_name,' Gene Frequency'),y = 'Gene Frequency Count',
         title = paste0(gene_name,' Gene Frequency Histogram'))+
    project_theme
  print(histogram)

  #Code to create scatter plot
  scatterplot <- ggplot(cleanedData,aes_string(x = gene_name,y = cont_covariate, colour = gene_name)) +
    geom_point()+
    #Creating gradient color representation
    scale_color_gradient(low = "blue", high = "red")+
    scale_x_continuous("Gene Frequency")+
    scale_y_continuous(paste0(cont_covariate))+
    labs(title = paste0('Gene Expression ', gene_name, ' vs. ', cont_covariate))+
    #Creating smooth trendline
    geom_smooth(aes(color=..y..), method = "loess", se = FALSE)+
    project_theme
  print(scatterplot)

  #Code to create boxplot
  boxplot <- ggplot(cleanedData,aes_string(x = cat_covariate1,y = gene_name,fill = cat_covariate2)) +
    # Add box plot
    geom_boxplot(color = "white") +
    scale_fill_manual(values = c("red", "purple"))+
    # Change labels
    labs(x = paste(cat_covariate1),y = paste0(gene_name, ' Gene Frequency'),fill =
           paste(cat_covariate2), title = paste0(gene_name, ' Gene Frequency by ',
                                cat_covariate1, ' and ', cat_covariate2))+
    project_theme
  print(boxplot)
}

#Calling the function with one gene example
# project_plots(df_name = combinedData, gene_name = 'AAAS', cont_covariate =
#'hospital.free_days_post_45_day_followup', cat_covariate1 = 'sex', cat_covariate2 = 'icu_status')

#Creating loop to go though multiple genes
multiple_gene_selection <- c('AAAS', 'AAGAB', 'AAMDC')
for (gene in multiple_gene_selection){
  project_plots(df_name = combinedData, gene_name = gene, cont_covariate =
    'hospital.free_days_post_45_day_followup', cat_covariate1 = 'sex', cat_covariate2 = 'icu_status')
}
```

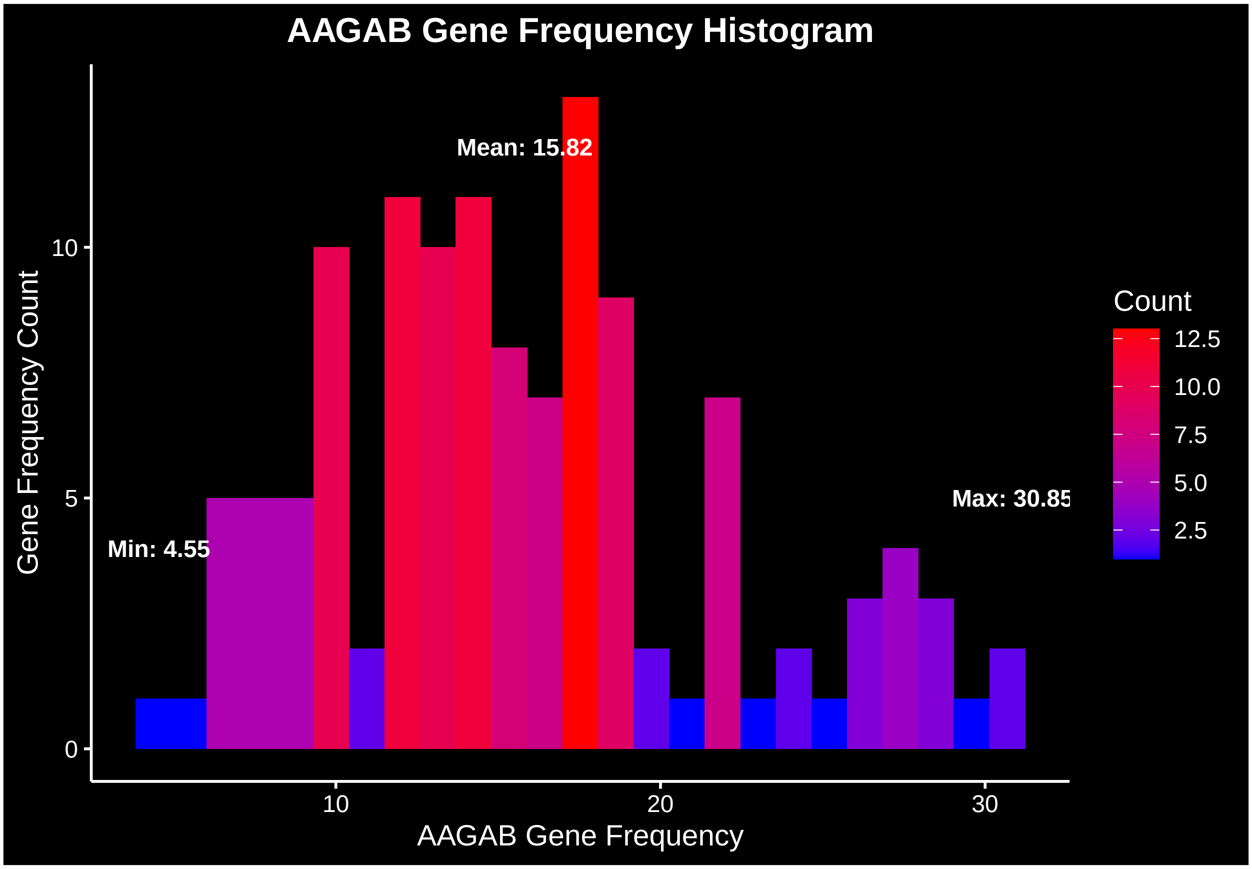
```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use ‘all_of()‘ or ‘any_of()‘ instead.
##   # Was:
##   data %>% select(gene_name)
##
##   # Now:
##   data %>% select(all_of(gene_name))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use ‘all_of()‘ or ‘any_of()‘ instead.
##   # Was:
##   data %>% select(cont_covariate)
##
##   # Now:
##   data %>% select(all_of(cont_covariate))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use ‘all_of()‘ or ‘any_of()‘ instead.
##   # Was:
##   data %>% select(cat_covariate1)
##
##   # Now:
##   data %>% select(all_of(cat_covariate1))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use ‘all_of()‘ or ‘any_of()‘ instead.
##   # Was:
##   data %>% select(cat_covariate2)
##
##   # Now:
##   data %>% select(all_of(cat_covariate2))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning: ‘aes_string()‘ was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with ‘aes()‘.
## i See also ‘vignette("ggplot2-in-packages")‘ for more information.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



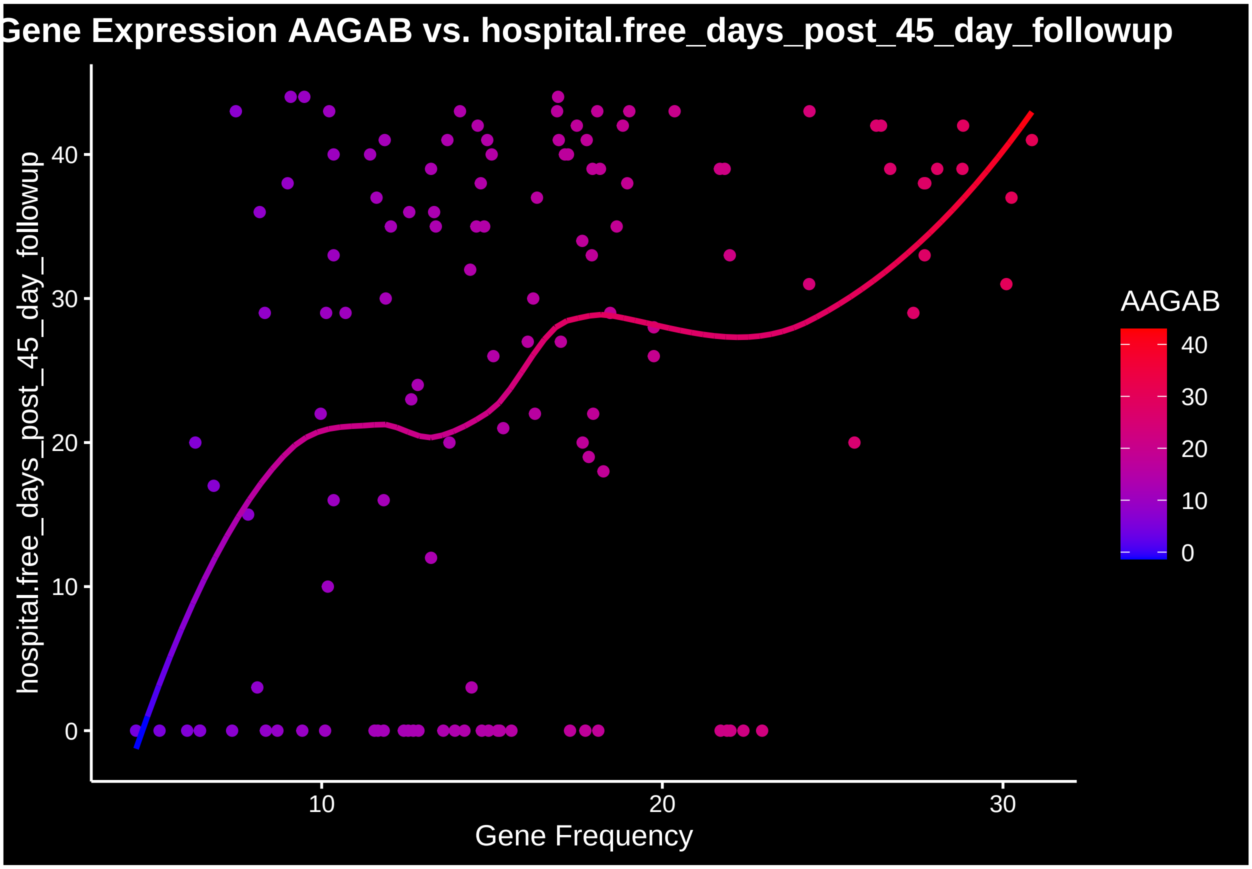
```
## `geom_smooth()` using formula = 'y ~ x'
```

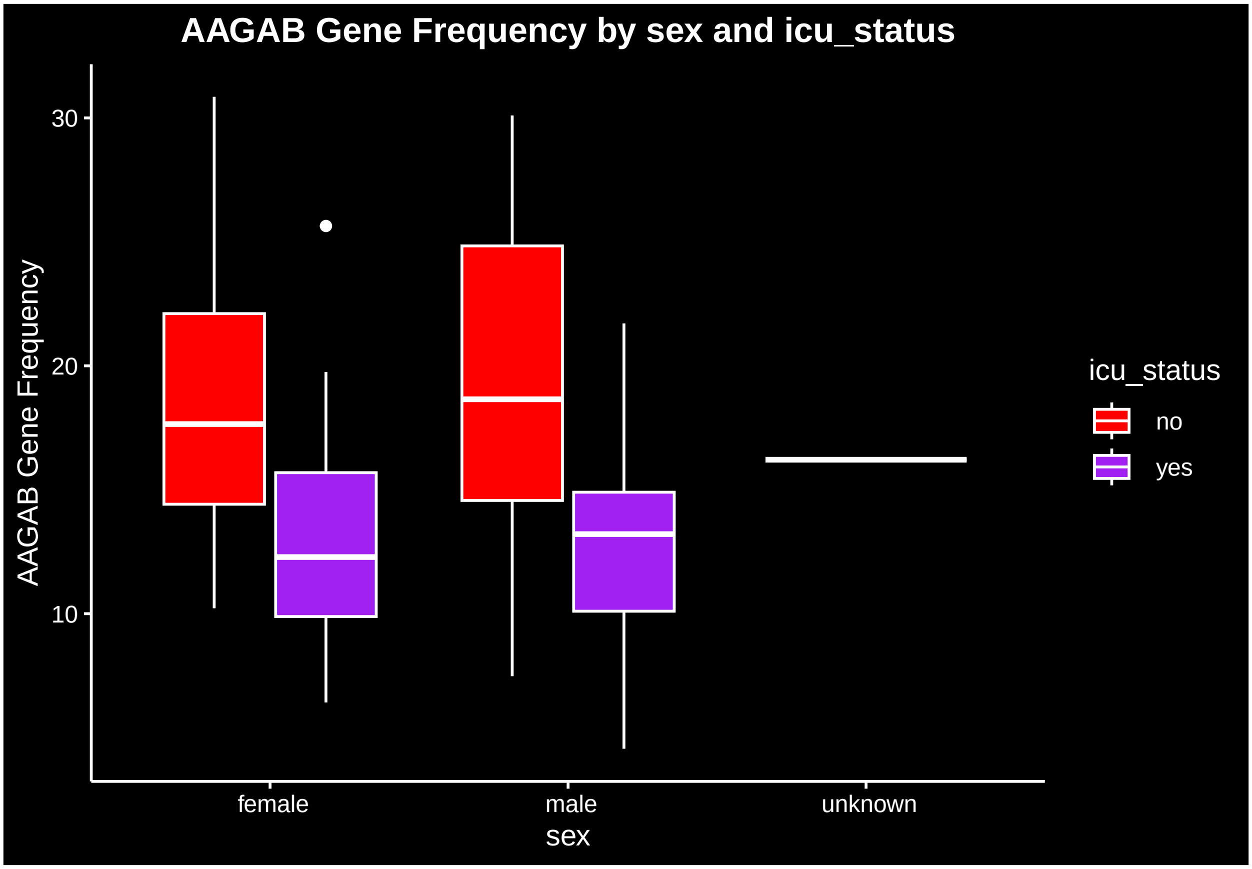

Gene Expression AAAS vs. hospital.free_days_post_45_day_followup
hospital.free_days_post_45_day_followup
0
10
20
30
40
10
20
30
Gene Frequency
AAAS
30
25
20
15
10

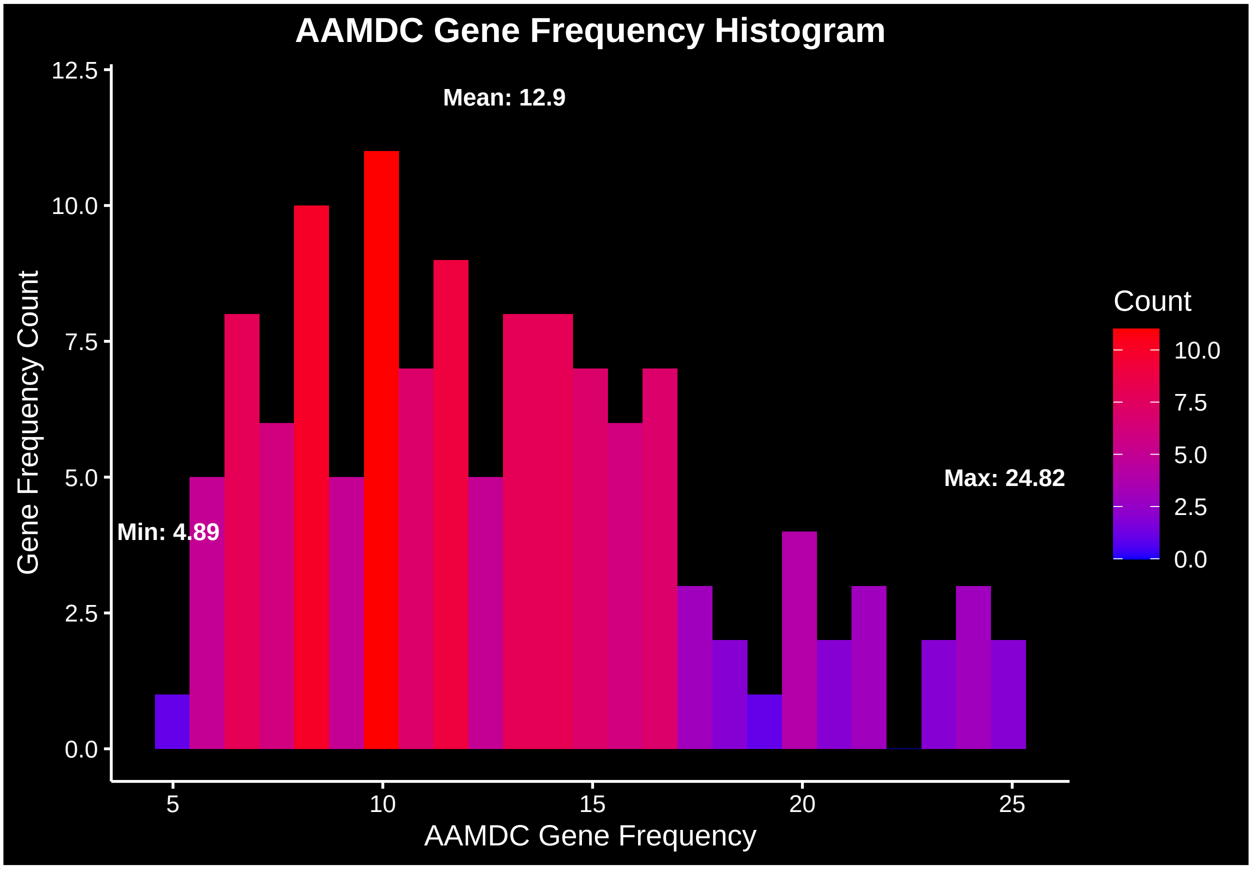

AAAS Gene Frequency by sex and icu_status
AAAS Gene Frequency
30
20
10
female
male
unknown
sex
icu_status
no
yes



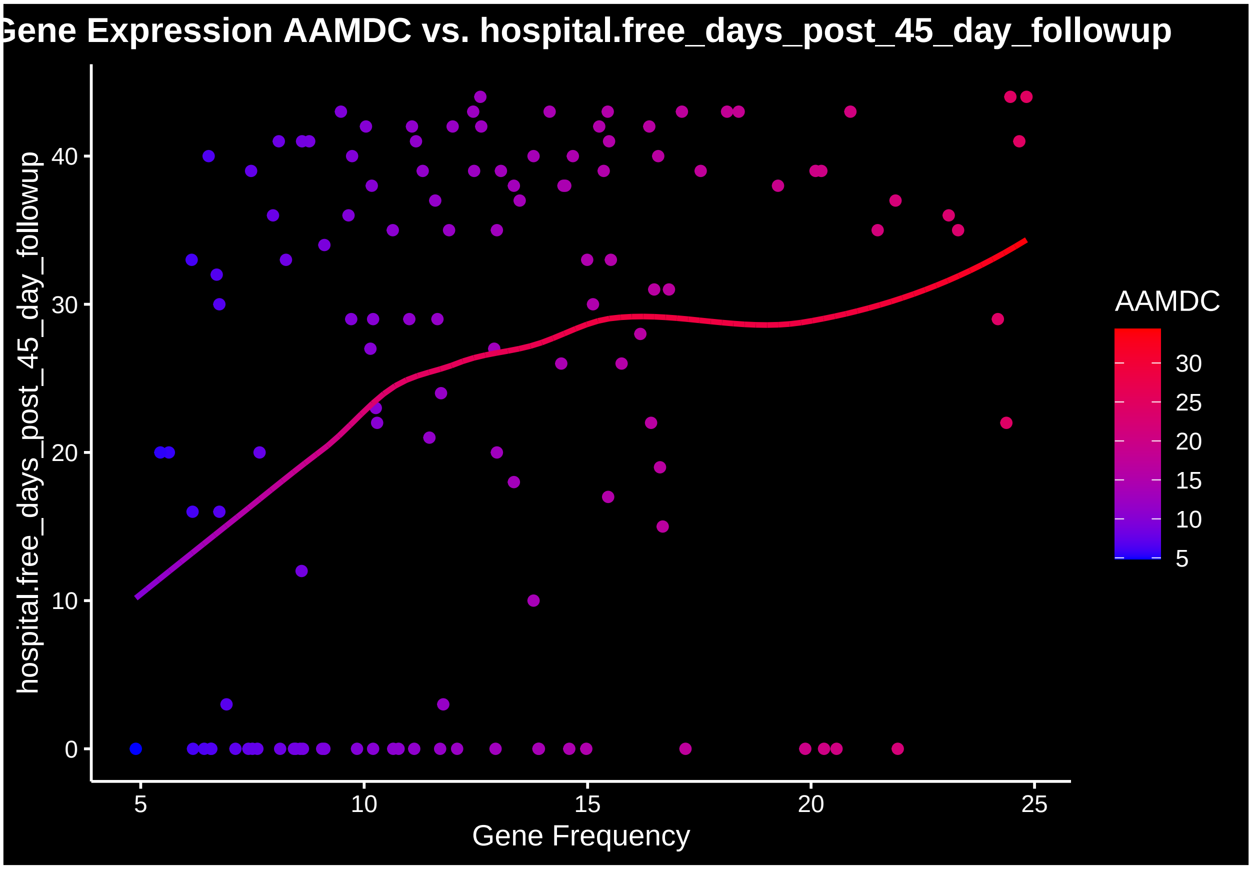
```
## 'geom_smooth()' using formula = 'y ~ x'
```

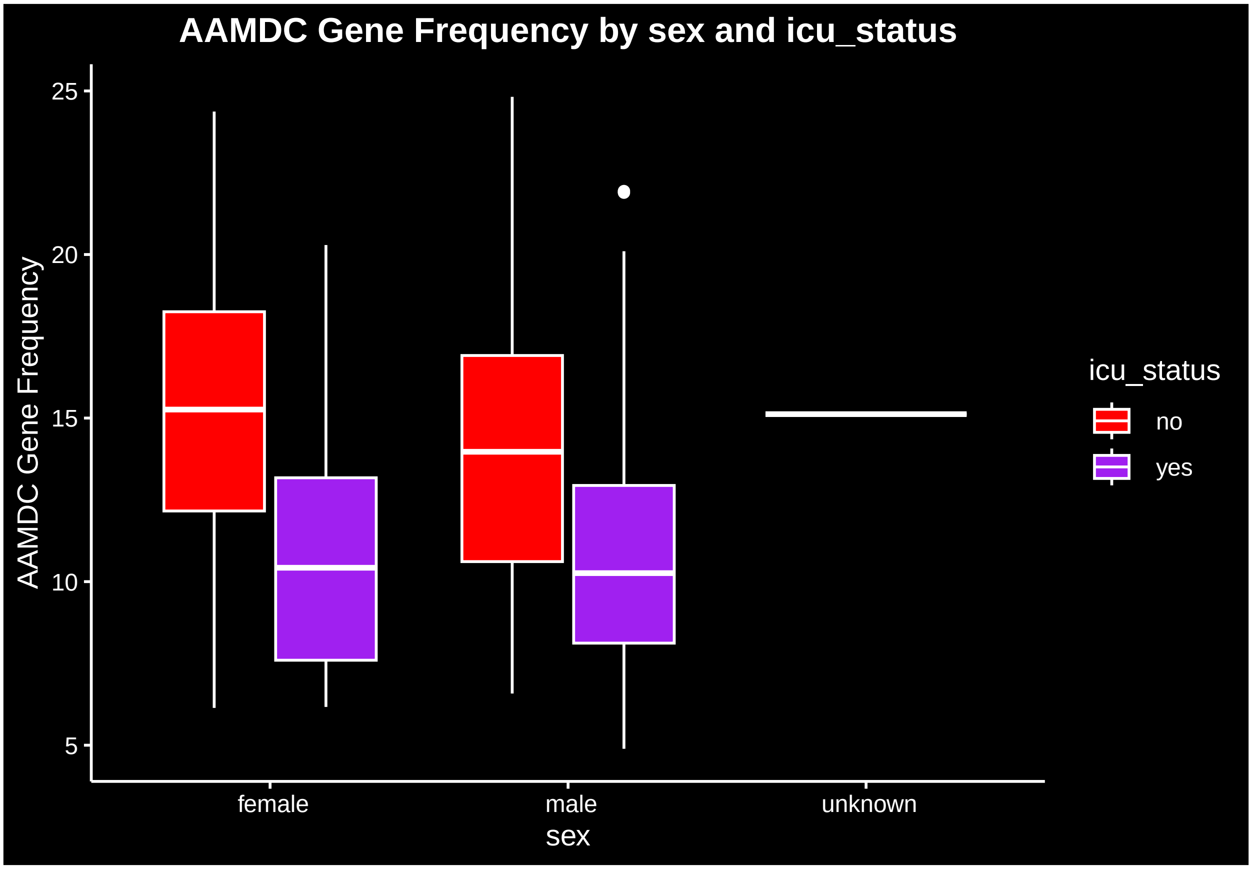

Gene Expression AAGAB vs. hospital.free_days_post_45_day_followup
hospital.free_days_post_45_day_followup
0
10
20
30
40
10
20
30
Gene Frequency
AAGAB
40
30
20
10
0


AAGAB Gene Frequency by sex and icu_status
AAGAB Gene Frequency
30
20
10
female
male
unknown
sex
icu_status
no
yes



```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```


Gene Expression AAMDC vs. hospital.free_days_post_45_day_followup
hospital.free_days_post_45_day_followup
0
10
20
30
40
5
10
15
20
25
Gene Frequency
AAMDC
30
25
20
15
10
5


AAMDC Gene Frequency by sex and icu_status
AAMDC Gene Frequency
25
20
15
10
5
female
male
unknown
sex
icu_status
no
yes