Project - Submission1

Rob Ross-Shannon

2024-07-25

```
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
```

```
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
```

```r
#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")

#Selecting gene of interest and transposing data to link with covariates
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)

#Fixing typo in data set
genesT$participant_id[which(genesT$participant_id=="COVID_06_.y_male_NonICU")] <- "COVID_06_:y_male_Non

#Selecting covariates of interest
covariates <- series_matrix[, c("participant_id","sex", "icu_status", "hospital.free_days_post_45_day_f

#Selecting gene of interest
gene_of_interest <- select(genesT, c("AAAS", "participant_id"))

#linking gene and covariate data set
combinedData <- inner_join(gene_of_interest, covariates, by = "participant_id")

#Removing any unknown values from sex dataset
cleanedData <- combinedData[combinedData$sex != " unknown",]
```

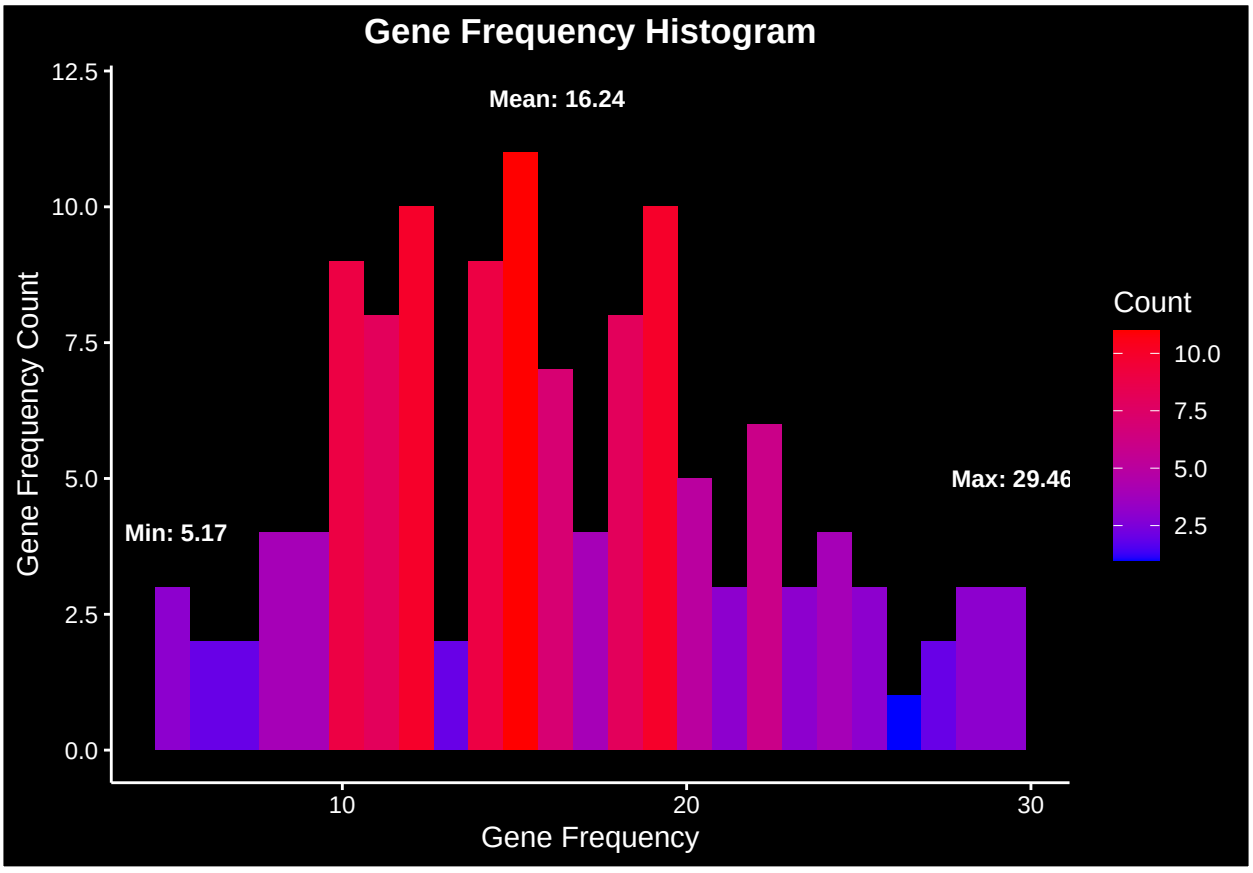
```r
#Establishing project theme for plots
project_theme <- theme(
        panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(hjust = 0.5, face = "bold"),
        # Define my axis
        title = element_text(colour = "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')
```

```r
#Creating annotations for plot labels
annotations <- data.frame(
  x = c(round(min(combinedData$AAAS), 2), round(mean(combinedData$AAAS), 2), round(max(combinedData$AAA
  y = c(4, 12, 5),
```

```
  label = c("Min:", "Mean:", "Max:"))

#Creating histogram
ggplot(combinedData,aes(x = AAAS)) +
  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"
  labs(x = 'Gene Frequency',y = 'Gene Frequency Count', title = "Gene Frequency Histogram")+
  project_theme
```

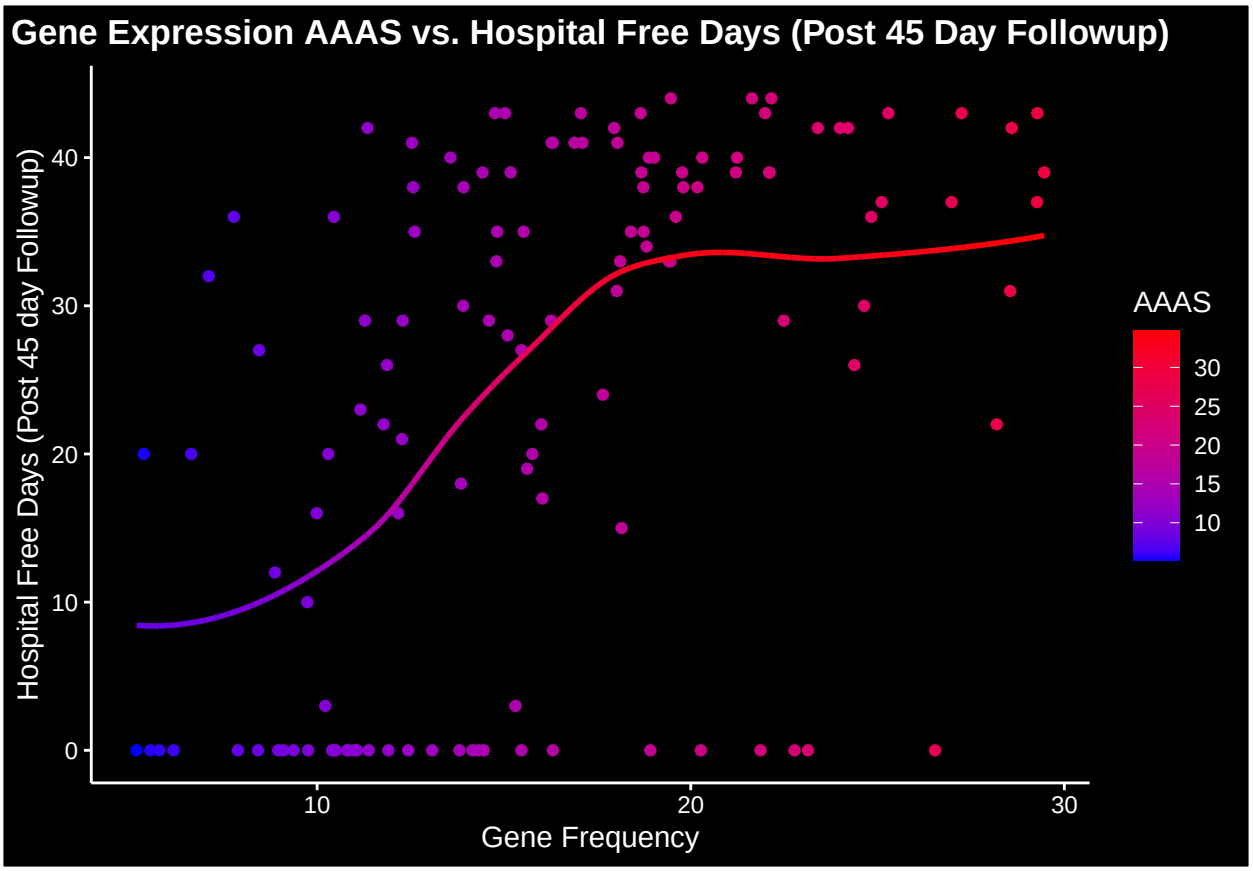
```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



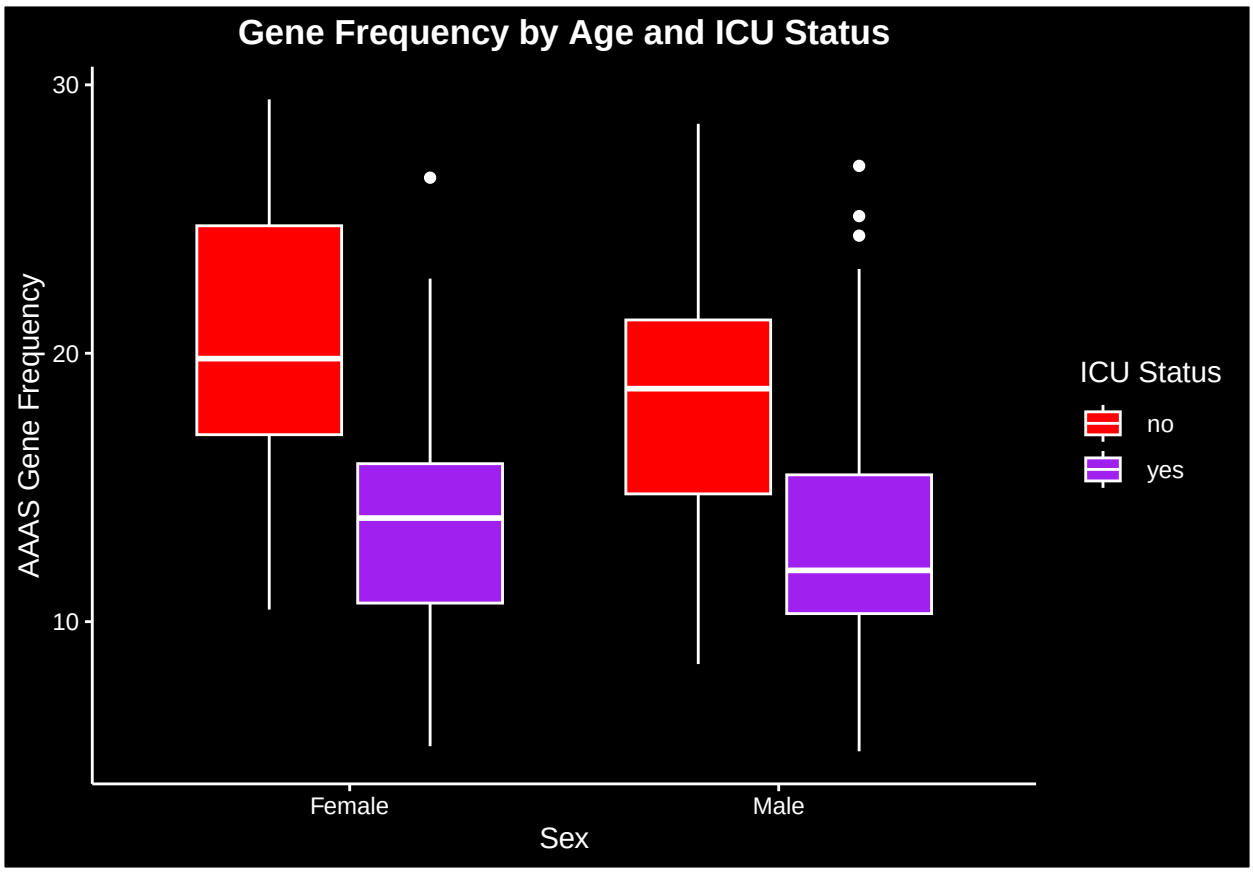

```
#Creating
ggplot(combinedData,aes(x = AAAS,y = hospital.free_days_post_45_day_followup, colour = AAAS)) +
  geom_point()+
  #Creating gradient color representation
  scale_color_gradient(low = "blue", high = "red")+
  scale_x_continuous("Gene Frequency")+
  scale_y_continuous("Hospital Free Days (Post 45 day Followup)")+
```

```
  labs(title = "Gene Expression AAAS vs. Hospital Free Days (Post 45 Day Followup)")+
  #Creating smooth trendline
  geom_smooth(aes(color=..y..), method = "loess", se = FALSE)+
  project_theme
```

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




```
ggplot(cleanedData,aes(x = sex,y = AAAS,fill = icu_status)) +
  # Add box plot
  geom_boxplot(color = "white") +
  scale_fill_manual(values = c("red", "purple"))+
  # Change labels
  labs(x = 'Sex',y = 'AAAS Gene Frequency',fill = 'ICU Status', title = "Gene Frequency by Age and ICU S
  scale_x_discrete(labels = c("Female", "Male"))+
  project_theme
```


Gene Frequency by Age and ICU Status
AAAS Gene Frequency
30
20
10
Female
Male
Sex
ICU Status
no
yes