QBS103_Submission 1

Angelique Cortez

2024-07-29

```
set.seed(1001001)
```

```
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

```
#read the downloaded csvs for the gene expression and metadata
gene_expression <- read.csv("~/Dropbox (Dartmouth College)/Dartmouth Classes/Summer2024/Intro to data an
metadata <- read.csv("~/Dropbox (Dartmouth College)/Dartmouth Classes/Summer2024/Intro to data analysis/
```

```
#view the data
#head(gene_expression)
```

```
#view data
#head(metadata)
```

```
#determine if file is a dataframe
class(gene_expression)
```

```
## [1] "data.frame"
```

```
class(metadata)
```

```
## [1] "data.frame"
```

#Identify one gene, one continuous covariate, and two categorical covariates in the provided dataset. Note: Gene expression data and metadata are in two separate files and will need to be linked.

```r
#need to set the gene name from X column
gene_expression <- gene_expression %>%
  rename(gene = X)
```

```r
library(tidyr)
library(tibble)
# Change the gene expression data to long format to combine later
gene_expression_long <- gene_expression %>%
  pivot_longer(cols = -gene, names_to = "participant_id", values_to = "expression")
```

```r
# Merge the datasets
combined_data <- merge(gene_expression_long, metadata, by.x = "participant_id", by.y = "participant_id"
```

##was running into issues and used this resources https://stackoverflow.com/questions/70191127/transforming-complete-age-from-character-to-numeric-in-r to mutate the age as numeric

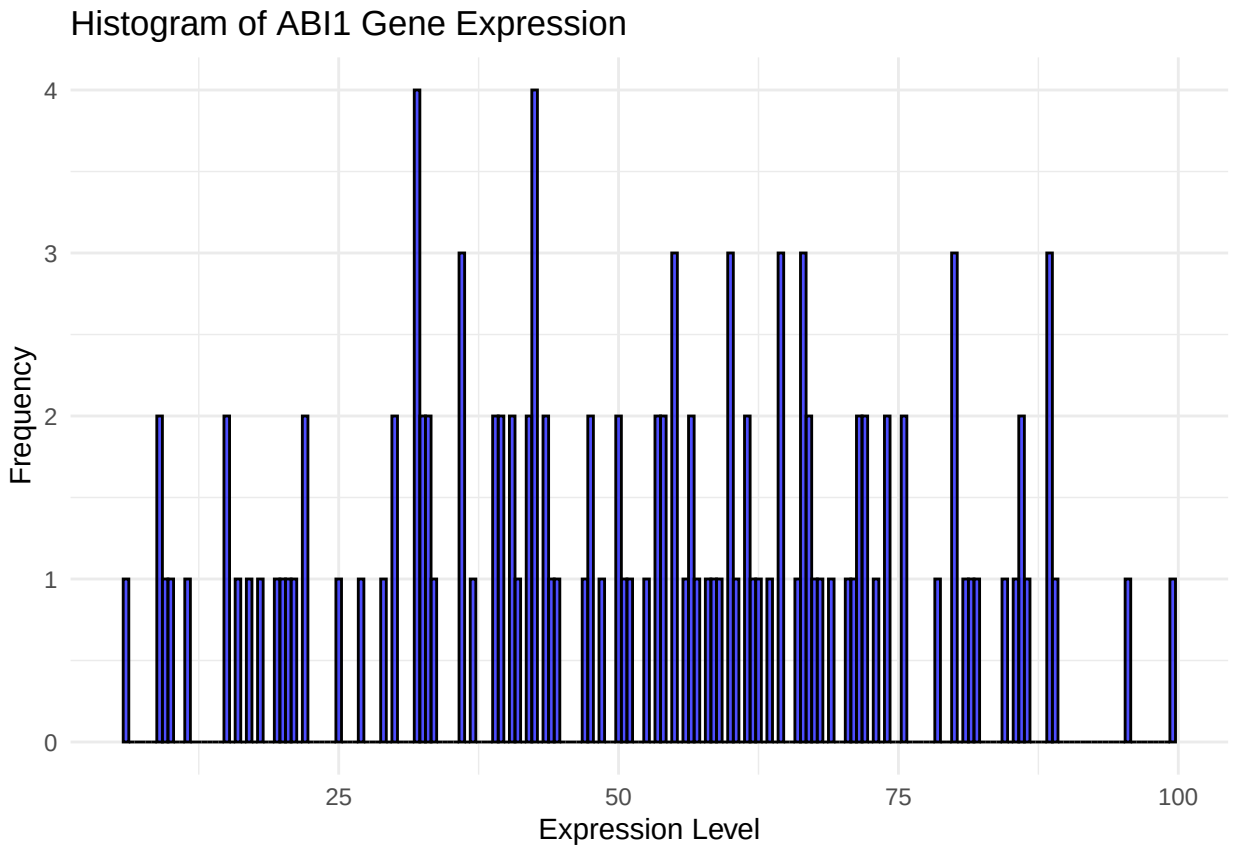
```r
# Choose gene of interest
gene_of_interest <- combined_data %>%
  filter(gene == "ABI1") %>% # choose 'ABI1' gene of interest
  mutate(
    # Extract age using a regular expression and convert to numeric
    age = as.numeric(sub(".*_(\\d{2})y_.*", "\\1", participant_id)),
    # Extract sex using a regular expression
    sex = sub(".*_(male|female)_.*", "\\1", participant_id)
  ) %>%
  select(participant_id, expression, age, sex, mechanical_ventilation) %>% #choose covariates
  column_to_rownames(var = "participant_id") # Set participant_id as row names
```

#realized I may want to have the correlating participant ID for later

#Generate the following three plots using ggplot2 for your covariates of choice: o Histogram for gene expression (5 pts)

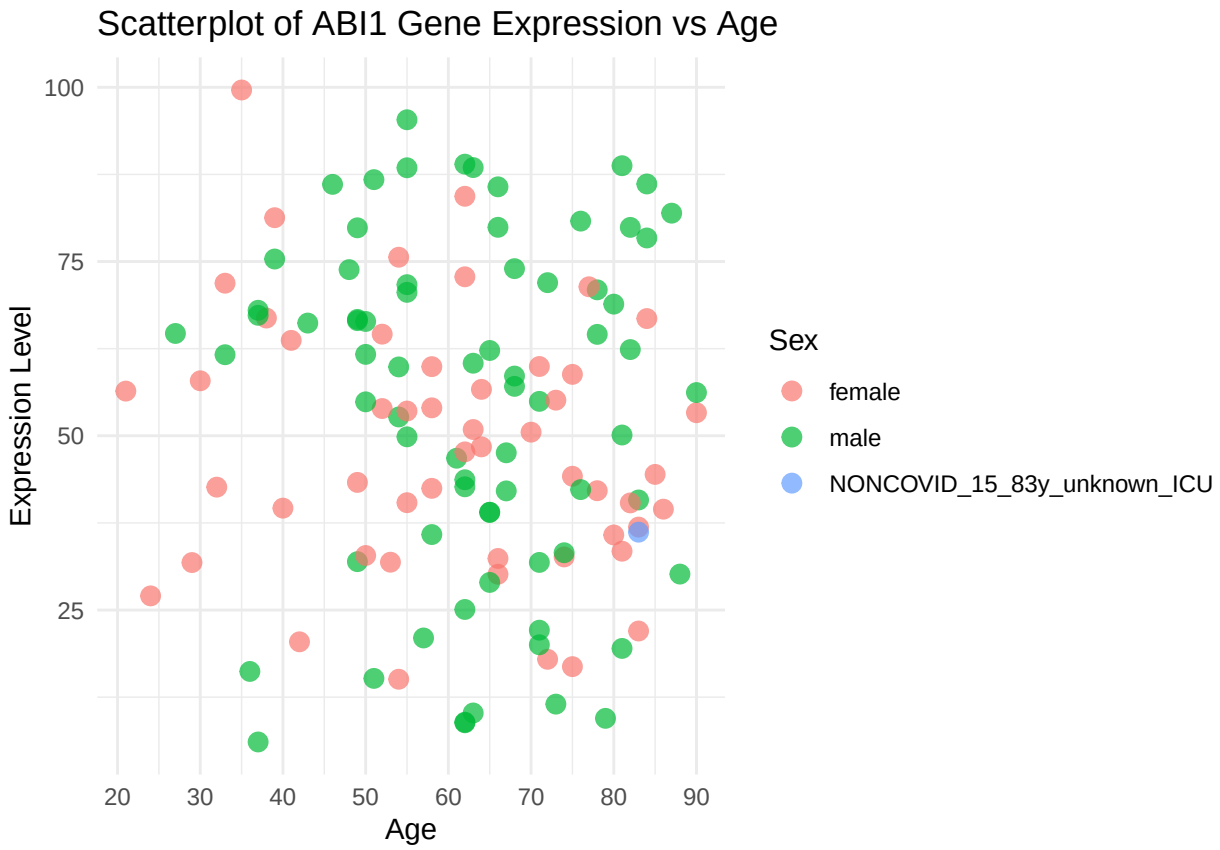
```r
#call library
library(ggplot2)
```

```r
# Create a histogram of the expression values
ggplot(gene_of_interest, aes(x = expression)) +
  geom_histogram(binwidth = 0.5, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of ABI1 Gene Expression",
       x = "Expression Level",
       y = "Frequency") +
  theme_minimal()
```



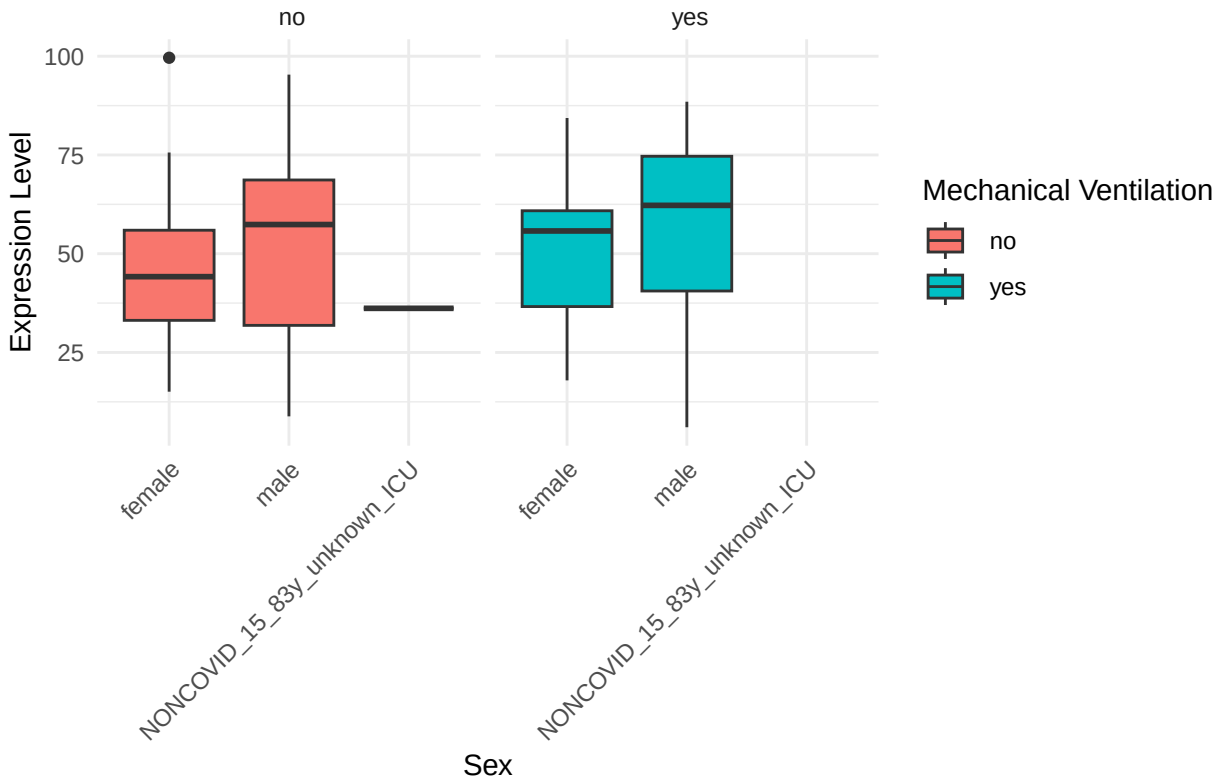
o Scatterplot for gene expression and continuous covariate (5 pts)

```
# Create a scatterplot of expression vs age, colored by sex
ggplot(gene_of_interest, aes(x = age, y = expression, color = sex)) +
  geom_point(size = 3, alpha = 0.7) +
  scale_x_continuous(breaks = seq(0, 100, by = 10)) +  # Customize x-axis breaks
  labs(title = "Scatterplot of ABI1 Gene Expression vs Age",
       x = "Age",
       y = "Expression Level",
       color = "Sex") +
  theme_minimal()
```



o Boxplot of gene expression separated by both categorical covariates (5 pts)

```
#Create a boxplot of expression separated by both sex and mechanical_ventilation
ggplot(gene_of_interest, aes(x = sex, y = expression, fill = mechanical_ventilation)) +
  geom_boxplot() +
  labs(title = "Boxplot of ABI1 Gene Expression by Sex and Mechanical Ventilation",
       x = "Sex",
       y = "Expression Level",
       fill = "Mechanical Ventilation") +
  theme_minimal() +
  facet_wrap(~ mechanical_ventilation)+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```

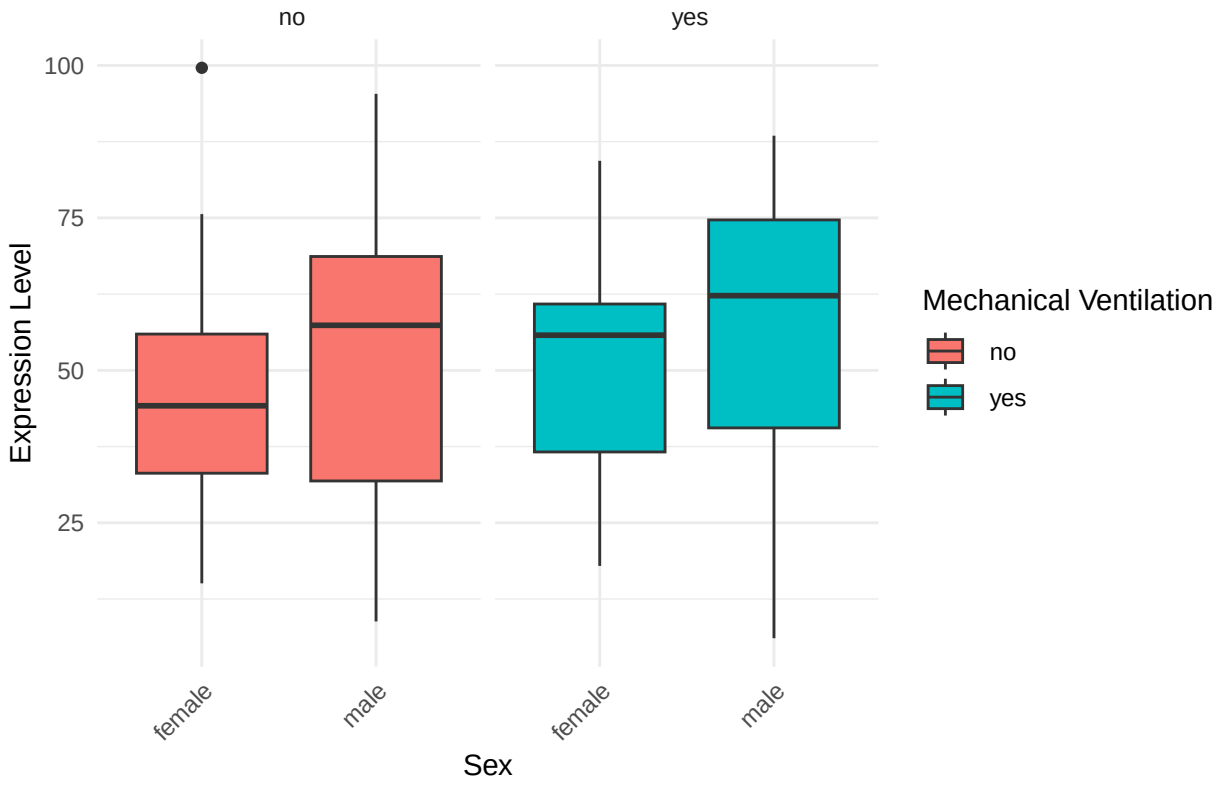


##Remove the unknown age https://www.r-bloggers.com/2022/06/remove-rows-from-the-data-frame-in-r/
https://www.tutorialspoint.com/how-to-remove-rows-in-an-r-data-frame-using-row-names

```
#remove unknown
gene_of_interest <- gene_of_interest[!(rownames(gene_of_interest) %in% "NONCOVID_15_83y_unknown_ICU"),
```

```
# Create a scatterplot of expression vs age, colored by sex
ggplot(gene_of_interest, aes(x = age, y = expression, color = sex)) +
  geom_point(size = 3, alpha = 0.7) +
  scale_x_continuous(breaks = seq(0, 100, by = 10)) +  # Customize x-axis breaks
  labs(title = "Scatterplot of ABI1 Gene Expression vs Age",
       x = "Age",
       y = "Expression Level",
       color = "Sex") +
  theme_minimal()
```


```
#Create a boxplot of expression separated by both sex and mechanical_ventilation
ggplot(gene_of_interest, aes(x = sex, y = expression, fill = mechanical_ventilation)) +
  geom_boxplot() +
  labs(title = "Boxplot of ABI1 Gene Expression by Sex and Mechanical Ventilation",
       x = "Sex",
       y = "Expression Level",
       fill = "Mechanical Ventilation") +
  theme_minimal() +
  facet_wrap(~ mechanical_ventilation)+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```


Boxplot of ABI1 Gene Expression by Sex and Mechanical Ventilation
no
yes
100
75
50
25
Expression Level
female
male
female
male
Sex
Mechanical Ventilation
no
yes