Project_part1_atreml

Alexandra Treml

2024-07-20

Part 1 includes: 1. set up git 2. read in the data from genes and series matrix 3. id 1 gene, 1 continuous variable, and 2 categorical covariates in the provided datasets (must link gene expression data and metadata) 4. Generate 3 plots using ggplot2 for your covariates of choice a. hist for gene expression b. scatterplot for gene expression and continuous covariate c. boxplot of gene expression separated by both categorical variables

#should filter out ICU from Non-ICU

R Markdown

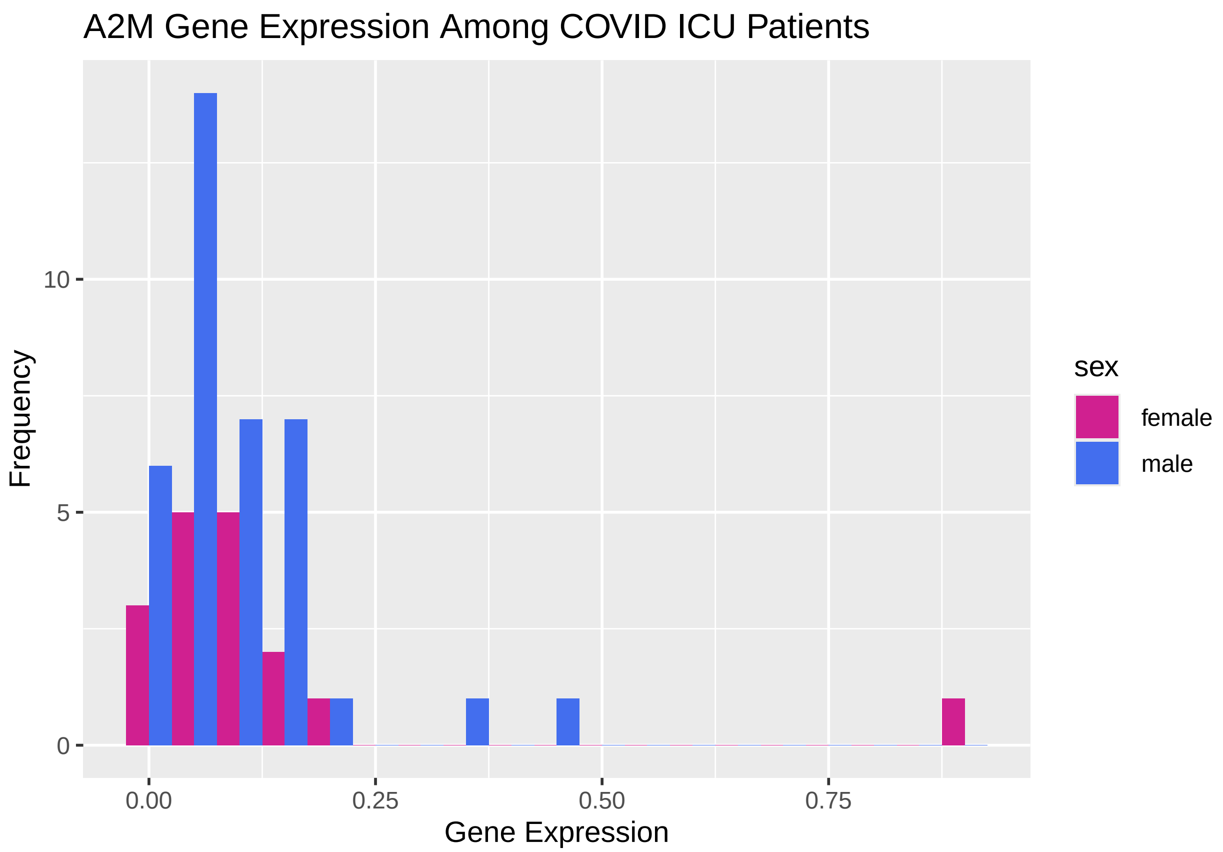
```r
# Set the working directory and read in the files
setwd("C:/Users/AlexandraTreml/Desktop/MS/QBS103/Project/QBS103_proj")
genes <- read.csv("genes_GSE157103.csv")
participant <- read.csv("series_matrix_GSE157103.csv")

#head(genes)
#head(participant)

# Rename the first column in genes to 'gene'
genes <- genes %>%
  rename(gene = X)

# Filter for 'A2M' gene
genes <- genes %>%
  filter(gene == "A2M")

# Pivot the genes dataframe
genes_long <- genes %>%
  tidyr::pivot_longer(cols = -gene, names_to = 'id', values_to = 'expression')
#head(genes_long)

# rename id column in participant, select my categorical covariates and continuous variable (+age for e
participant <- participant %>%
  filter(icu_status == ' yes') %>%
  rename(id = participant_id) %>%
  select(id, age, sex, mechanical_ventilation, lactate.mmol.l.)
#head(participant)

#join 2 dataframes
df <- left_join(participant, genes_long, by = "id")
```

```r
# head(df)

#clean up new dataframe

#fill lactate unknowns with NA
df$lactate.mmol.l.[df$lactate.mmol.l. == " unknown"] <- NA
#fill sex unknowns with NA
df$sex[df$sex == " unknown"] <- NA
#remove : from age, and fill with NA
df$age[df$age == " :"] <- NA

#head(df)
```

```r
#create a histogram using sex and gene expression

# Load the necessary library
library(ggplot2)
df <- na.omit(df)
df <- df %>%
  select(id, gene, sex, age, mechanical_ventilation, expression, lactate.mmol.l.)
#head(df)

df$sex <- as.factor(df$sex)
levels(df$sex)
```

```
## [1] " female" " male"
```

```r
ggplot(data = df, aes(x = expression, fill = sex)) +
  geom_histogram(binwidth = 0.05, position = 'dodge') +
  labs(x = 'Gene Expression', y = 'Frequency', title = 'A2M Gene Expression Among COVID ICU Patients') +
  scale_fill_manual(values = c("violetred", "royalblue2"))
```



```r
#create a scatterplot using gene expression and lactate

# Load necessary libraries
library(dplyr)
library(ggplot2)

#head(df)

# Convert lactate.mmol.l. to numeric
df$lactate.mmol.l. <- as.numeric(df$lactate.mmol.l.)

# Remove leading and trailing whitespaces from sex column
df$sex <- trimws(df$sex)
df$mechanical_ventilation <- trimws(df$mechanical_ventilation)

# Check unique levels
unique(df$sex)
```

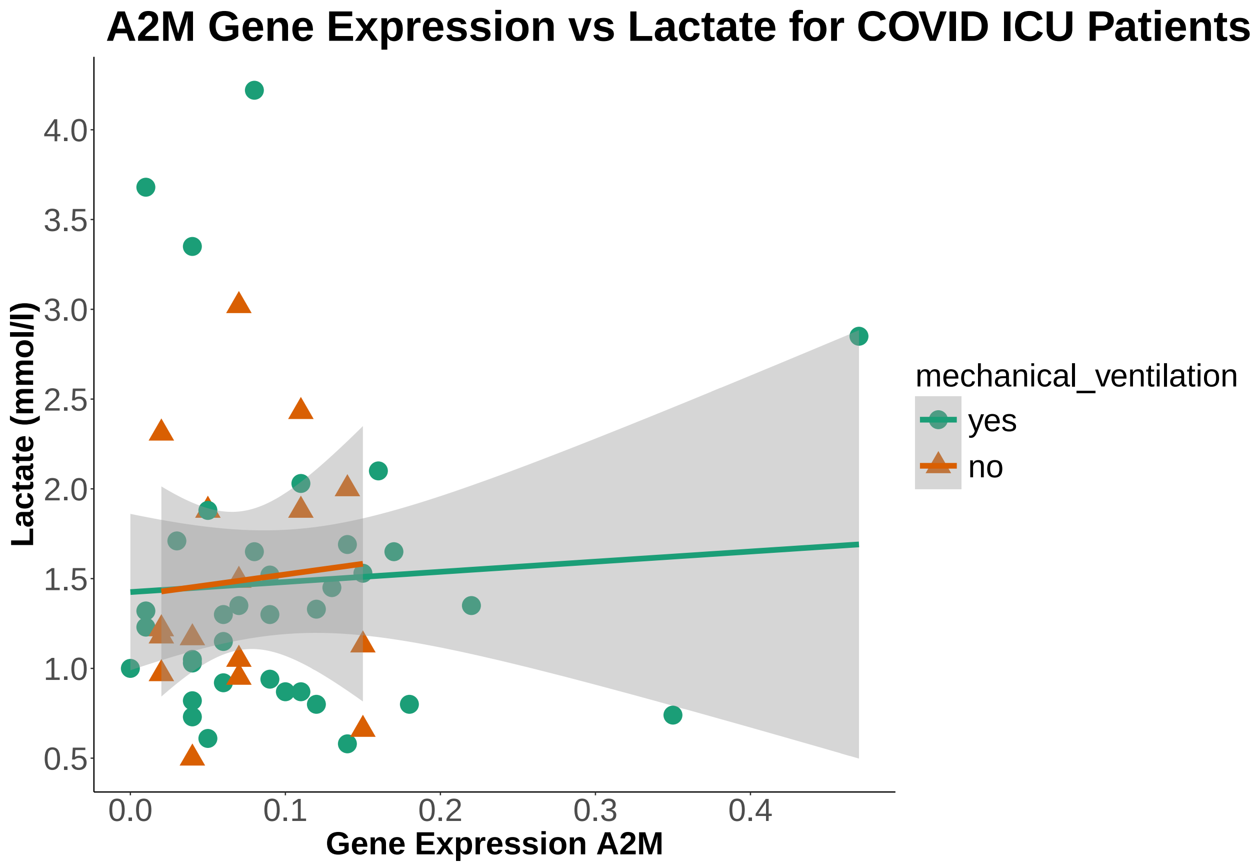
```
## [1] "male"   "female"
```

```r
unique(df$mechanical_ventilation)
```

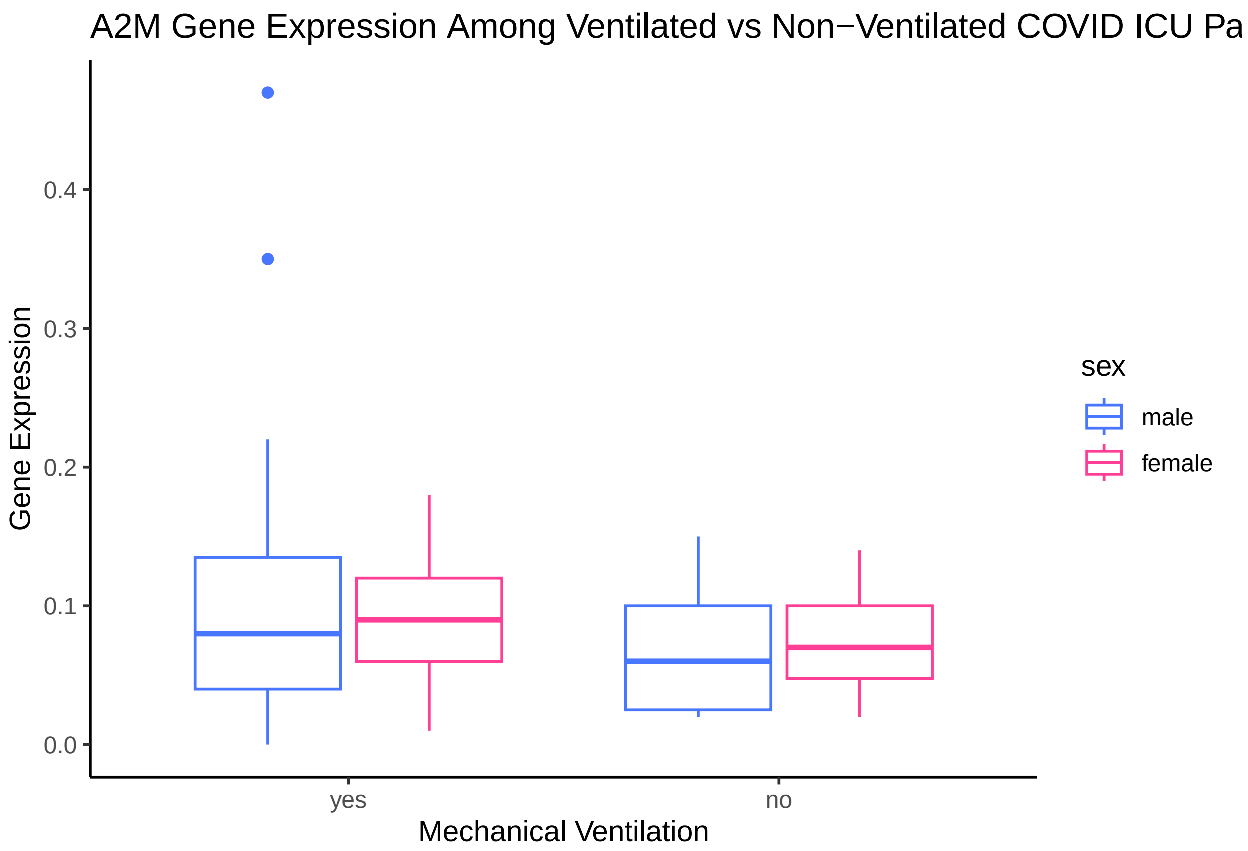
```
## [1] "yes" "no"
```

```
# Set factor levels
df$sex <- factor(df$sex, levels = c("male", "female"))
df$mechanical_ventilation <- factor(df$mechanical_ventilation, levels = c("yes", "no"))


# Filter out values greater than 6.5 for the y-axis
df <- df %>% filter(lactate.mmol.l. <= 6.5)

df <- df %>% filter(expression <= 0.5)

# scatterplot vent
ggplot(df, aes(x = expression, y = lactate.mmol.l., color = mechanical_ventilation, shape = mechanical_
  geom_point(size = 3) +
  scale_y_continuous(breaks = seq(0, ceiling(max(df$lactate.mmol.l., na.rm = TRUE)), by = 0.5)) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(x = 'Gene Expression A2M', y = 'Lactate (mmol/l)', title = ' A2M Gene Expression vs Lactate for
  theme_classic(base_size = 5) +
  scale_color_brewer(palette = 'Dark2') +
  #scale_color_manual(values = c("male" = "royalblue2", "female" = "violetred3")) +
  scale_shape_manual(values = c(16, 17)) +
  theme(
    plot.title = element_text(size = 16, face = "bold"), # Center and bold title
    axis.title = element_text(size = 12, face = "bold"), # Bold axis titles
    axis.text = element_text(size = 12), # Increase axis text size
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 12)
  )
```

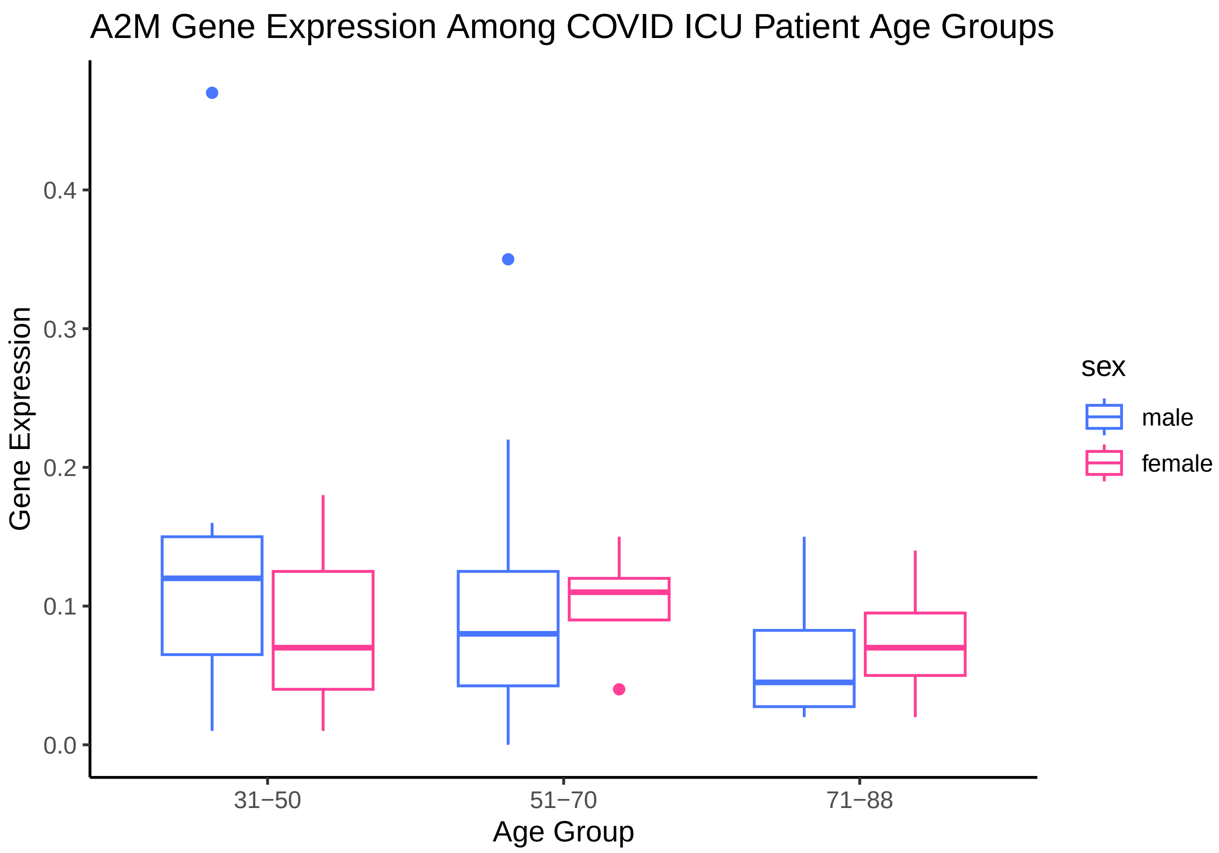
```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
#boxplot of gene expression separated by both ventilator, sex, and age group

#head(df)

df$age <- as.numeric(df$age)

df$age_group <- cut(df$age,
                    breaks = c(21, 30, 50, 70, 88),
                    labels = c("21-30", "31-50", "51-70", "71-88"),
                    right = TRUE)
#head(df)


ggplot(df, aes(x = mechanical_ventilation, y = expression, color = sex)) +
  geom_boxplot() +
  scale_color_manual(values = c("male" = "royalblue1", "female" = "violetred1")) +
  labs(x = 'Mechanical Ventilation', y = 'Gene Expression', title = 'A2M Gene Expression Among Ventilat
  theme_classic()
```



```
ggplot(df, aes(x = age_group, y = expression, color = sex)) +
  geom_boxplot() +
  scale_color_manual(values = c("male" = "royalblue1", "female" = "violetred1")) +
  labs(x = 'Age Group', y = 'Gene Expression', title = 'A2M Gene Expression Among COVID ICU Patient Age
  theme_classic()
```


A2M Gene Expression Among COVID ICU Patient Age Groups
Gene Expression
0.0
0.1
0.2
0.3
0.4
31−50
51−70
71−88
Age Group
sex
male
female