Project_part3

Alexandra Treml

2024-08-25

DATA CLEANING AND PREP

```
#DATA CLEANING AND PREP

# 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")

# Rename the first column in genes to 'gene'
genes <- genes %>%
  rename(gene = X) %>%
  filter(gene == "A2M")

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

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

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

#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

df <- na.omit(df) #get rid of na
```

Part 1: Latex table of summary stats for all variables.

```
# Load necessary libraries
library(dplyr)
library(stargazer)
```

```r
#library(knitr)

# Convert columns to appropriate types
df$fibrinogen <- as.numeric(df$fibrinogen)
```

```
## Warning: NAs introduced by coercion
```

```r
df$lactate.mmol.l. <- as.numeric(df$lactate.mmol.l.)
df$expression <- as.numeric(df$expression)
df$sex <- as.factor(df$sex)
df$mechanical_ventilation <- as.factor(df$mechanical_ventilation)
df$gene <- as.factor(df$gene)
df$crp.mg.l. <- as.numeric(df$crp.mg.l.)
```

```
## Warning: NAs introduced by coercion
```

```r
# Summary of numeric variables
stargazer(df %>% select(expression, lactate.mmol.l., fibrinogen, crp.mg.l.),
          type = "text",
          summary.stat = c("mean", "sd", "n"),
          digits = 2)
```

```
##
## ===================================
## Statistic        Mean  St. Dev. N
## -----------------------------------
## expression       0.11    0.13   54
## lactate.mmol.l.  1.64    1.39   54
## fibrinogen      501.82  216.97  45
## crp.mg.l.       144.71  102.81  48
## -----------------------------------
```

```r
sex_counts <- df %>%
  group_by(sex) %>%
  summarize(
    count = n(),
    n_percentage = (n()/nrow(df)) *100)

gene_counts <- df %>%
  group_by(gene) %>%
  summarize(
    count = n(),
    n_percentage = (n()/nrow(df)) *100)

vent_counts <- df %>%
  group_by(mechanical_ventilation) %>%
  summarize(
    count = n(),
    n_percentage = (n()/nrow(df)) *100)

sex_counts <- sex_counts %>%
```

```
  mutate(variable = "Sex")

gene_counts <- gene_counts %>%
  mutate(variable = "Gene")

vent_counts <- vent_counts %>%
  mutate(variable = "Mechanical Ventilation")

cat_variables <- bind_rows(sex_counts, gene_counts, vent_counts)

cat_variables <- cat_variables%>%
  select(variable, count, n_percentage)

stargazer(cat_variables, type = "text", summary = FALSE, rownames = FALSE)
```

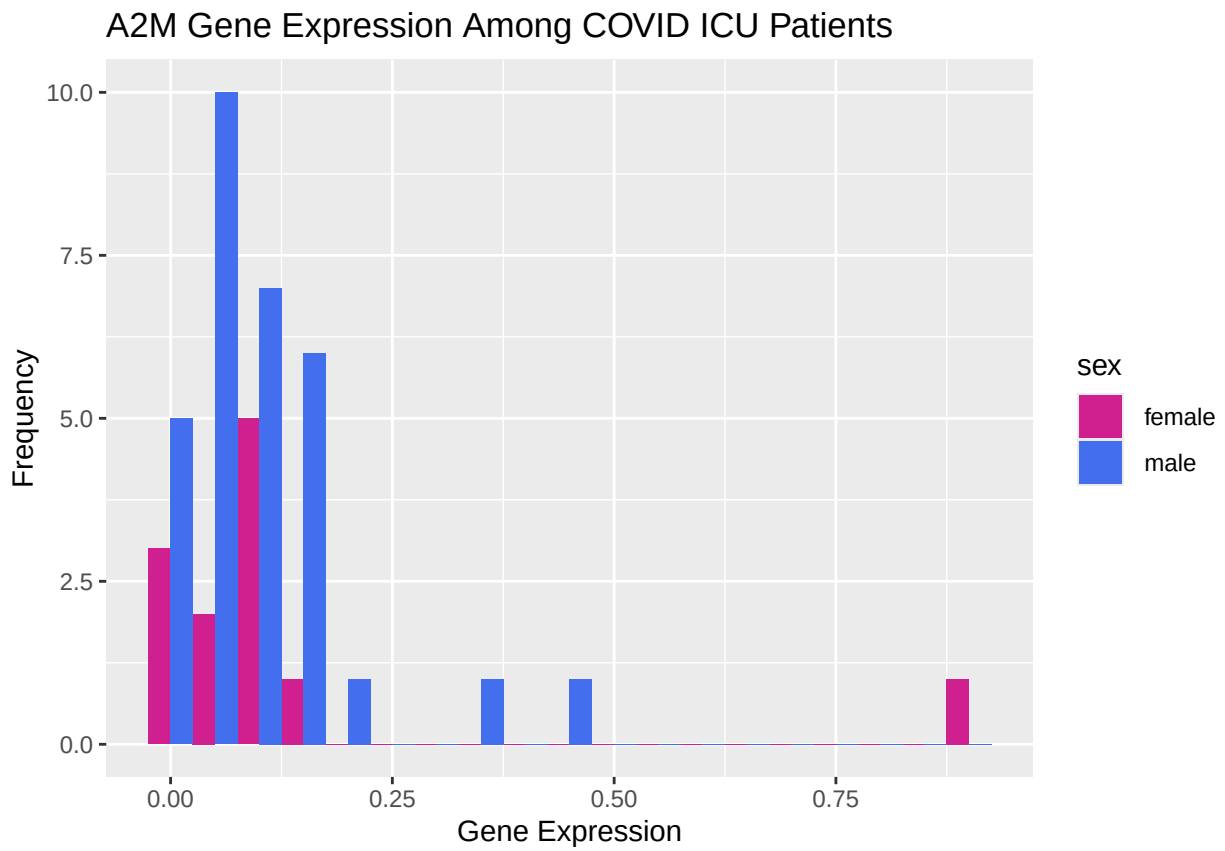
```
##
## ============================================
## variable               count   n_percentage
## --------------------------------------------
## Sex                     17   31.4814814814815
## Sex                     37   68.5185185185185
## Gene                    54         100
## Mechanical Ventilation  17   31.4814814814815
## Mechanical Ventilation  37   68.5185185185185
## --------------------------------------------
```

Part 2: Generate a publication quality histogram, boxplot, and scatterplot

```
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"
```

```
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"))
```




```
# 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"
```

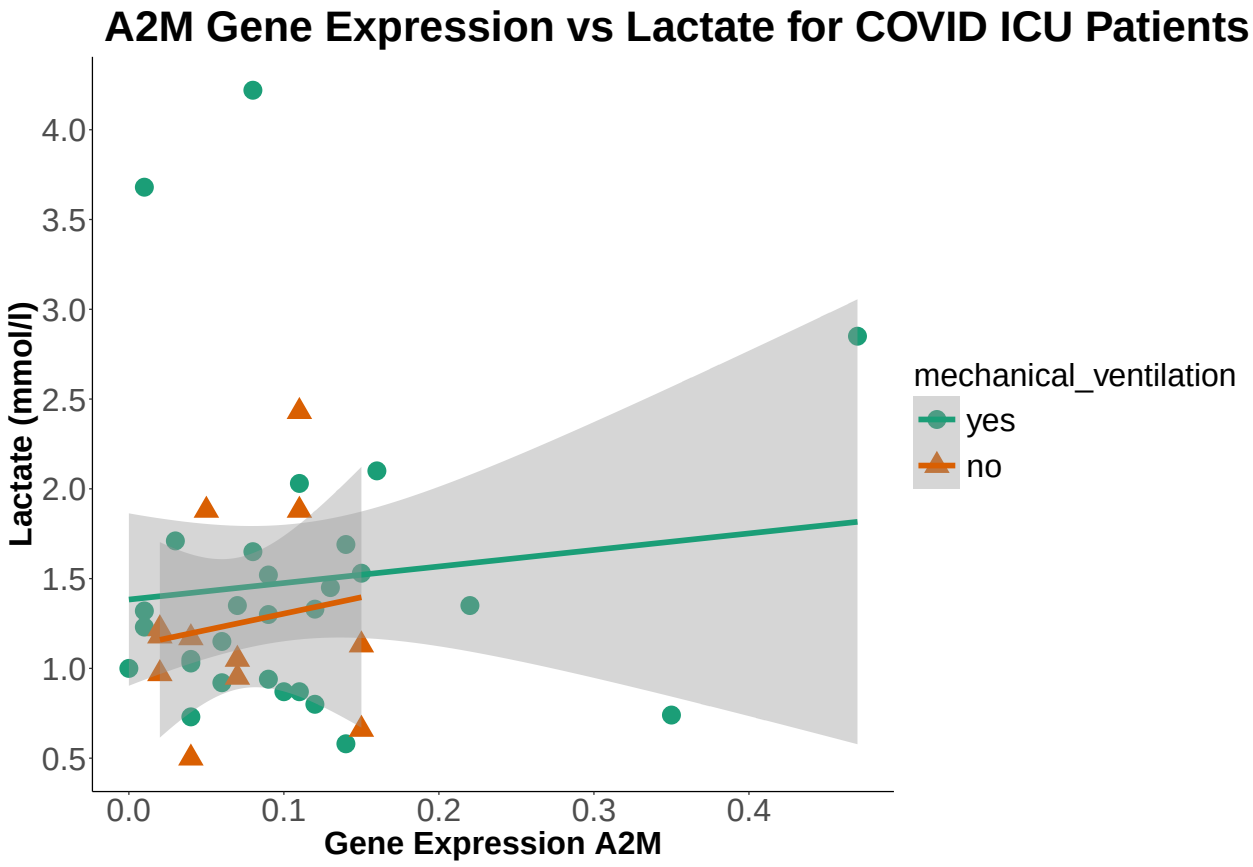
```
unique(df$mechanical_ventilation)
```

```
## [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)
```

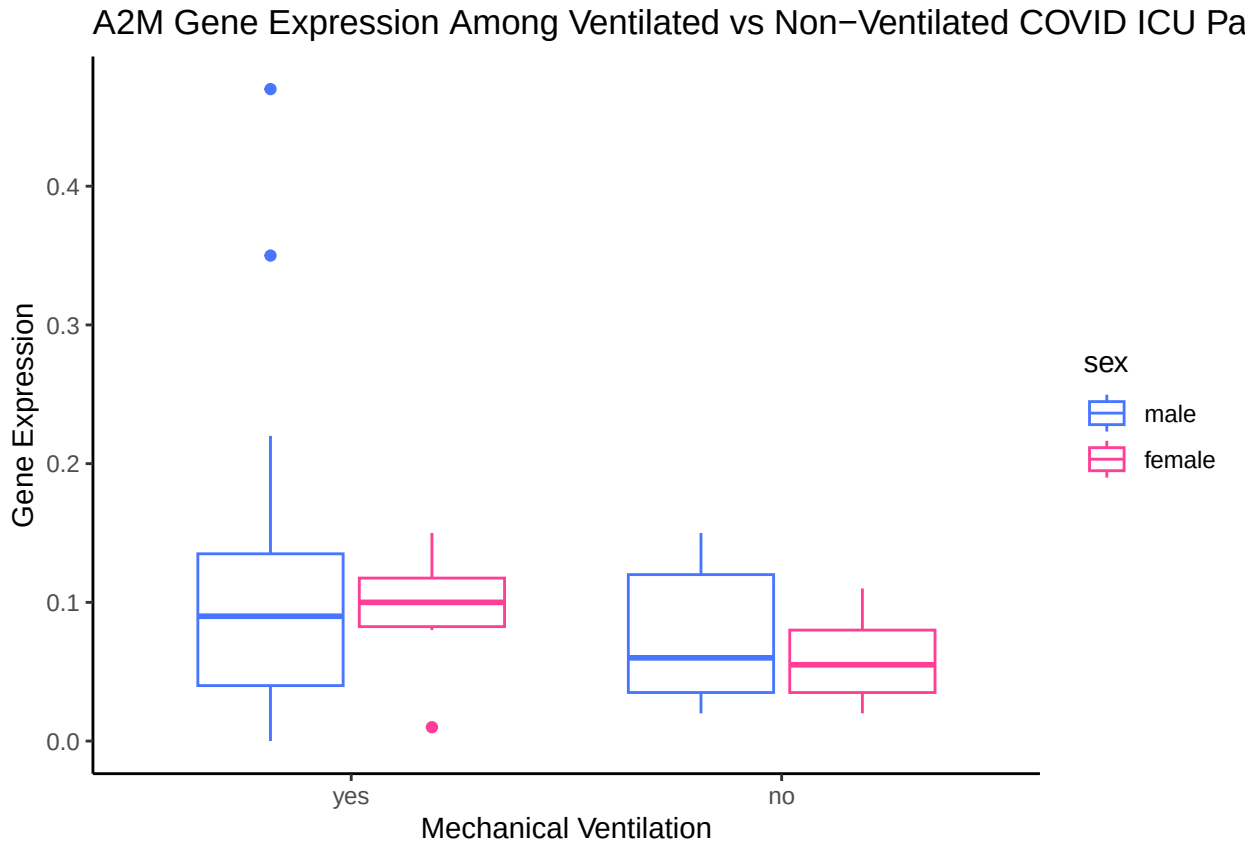
```r
df <- df %>% filter(expression <= 0.5)

# scatterplot vent
ggplot(df, aes(x = expression, y = lactate.mmol.l., color = mechanical_ventilation, shape = mechanical_v
  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 C
  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'
```




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

```r
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()
```



DATA PREP

```r
genes_1 <- read.csv("genes_GSE157103.csv")
participant <- read.csv("series_matrix_GSE157103.csv")


genes_1 <- genes_1 %>%
  rename(gene = X)

# Pivot the genes dataframe
genes_long_1 <- genes_1 %>%
  tidyr::pivot_longer(cols = -gene, names_to = 'id', values_to = 'expression')
```

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

#join 2 dataframes
df1 <- left_join(participant, genes_long_1, by = "id")

#fill lactate unknowns with NA
df1$lactate.mmol.l.[df1$lactate.mmol.l. == " unknown"] <- NA
#fill sex unknowns with NA
df1$sex[df1$sex == " unknown"] <- NA
#remove : from age, and fill with NA
df1$age[df1$age == " :"] <- NA

df1 <- na.omit(df1) #get rid of na

head(df1)
```

```
##                     id geo_accession                status
## 1 COVID_08_78y_male_ICU    GSM4753028 Public on Aug 29 2020
## 2 COVID_08_78y_male_ICU    GSM4753028 Public on Aug 29 2020
## 3 COVID_08_78y_male_ICU    GSM4753028 Public on Aug 29 2020
## 4 COVID_08_78y_male_ICU    GSM4753028 Public on Aug 29 2020
## 5 COVID_08_78y_male_ICU    GSM4753028 Public on Aug 29 2020
## 6 COVID_08_78y_male_ICU    GSM4753028 Public on Aug 29 2020
##   X.Sample_submission_date last_update_date type channel_count
## 1              Aug 28 2020      Aug 29 2020  SRA             1
## 2              Aug 28 2020      Aug 29 2020  SRA             1
## 3              Aug 28 2020      Aug 29 2020  SRA             1
## 4              Aug 28 2020      Aug 29 2020  SRA             1
## 5              Aug 28 2020      Aug 29 2020  SRA             1
## 6              Aug 28 2020      Aug 29 2020  SRA             1
##               source_name_ch1 organism_ch1          disease_status age   sex
## 1 Leukocytes from whole blood Homo sapiens disease state: COVID-19  78  male
## 2 Leukocytes from whole blood Homo sapiens disease state: COVID-19  78  male
## 3 Leukocytes from whole blood Homo sapiens disease state: COVID-19  78  male
## 4 Leukocytes from whole blood Homo sapiens disease state: COVID-19  78  male
## 5 Leukocytes from whole blood Homo sapiens disease state: COVID-19  78  male
## 6 Leukocytes from whole blood Homo sapiens disease state: COVID-19  78  male
##   icu_status apacheii charlson_score mechanical_ventilation
## 1        yes       43              7                    yes
## 2        yes       43              7                    yes
## 3        yes       43              7                    yes
## 4        yes       43              7                    yes
## 5        yes       43              7                    yes
## 6        yes       43              7                    yes
##   ventilator.free_days hospital.free_days_post_45_day_followup ferritin.ng.ml.
## 1                    0                                       0            1103
## 2                    0                                       0            1103
## 3                    0                                       0            1103
## 4                    0                                       0            1103
```

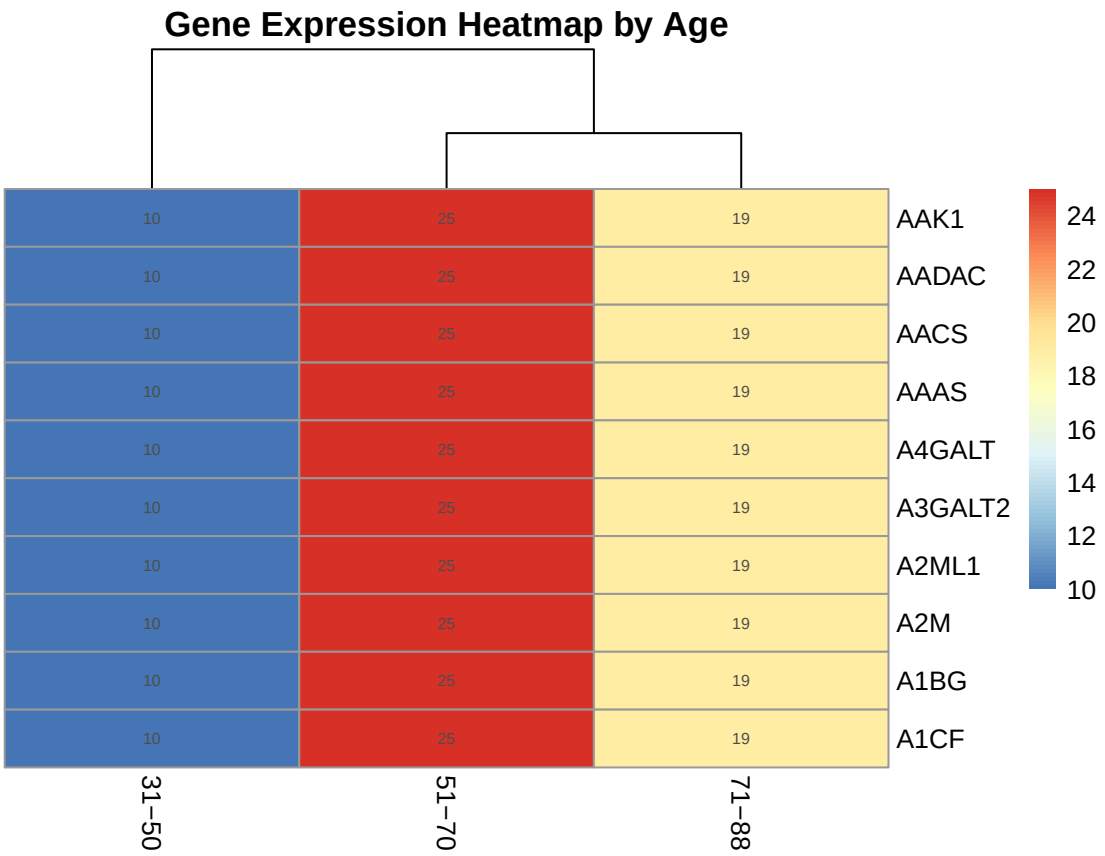
```
## 5                   0                                       0           1103
## 6                   0                                       0           1103
##   crp.mg.l. ddimer.mg.l_feu. procalcitonin.ng.ml.. lactate.mmol.l. fibrinogen
## 1      79.5           12.16                   4.2           1.65        780
## 2      79.5           12.16                   4.2           1.65        780
## 3      79.5           12.16                   4.2           1.65        780
## 4      79.5           12.16                   4.2           1.65        780
## 5      79.5           12.16                   4.2           1.65        780
## 6      79.5           12.16                   4.2           1.65        780
##   sofa    gene expression
## 1   16    A1BG       0.12
## 2   16    A1CF       0.00
## 3   16     A2M       0.08
## 4   16   A2ML1       0.01
## 5   16 A3GALT2       0.00
## 6   16  A4GALT       0.00
```

Part 3: Generate a Heat Map

```r
library(reshape2)
library(pheatmap)

ten_genes <- c("A1BG", "A1CF", "A2M", "A2ML1", "A3GALT2", "A4GALT", "AAAS", "AACS", "AADAC", "AAK1")

df1$age <- as.numeric(df1$age)

df1$age_group <- cut(df1$age,
                     breaks = c(21, 30, 50, 70, 88),
                     labels = c("21-30", "31-50", "51-70", "71-88"),
                     right = TRUE)

df_hm <- df1 %>%
  select(age_group, gene, expression, mechanical_ventilation, sex)%>%
  filter(gene %in% ten_genes)

#reshape
heatmap_data <- dcast(df_hm, gene ~ age_group, value.var = "expression")
```

```
## Aggregation function missing: defaulting to length
```

```r
rownames(heatmap_data) <- heatmap_data$gene
heatmap_data <- heatmap_data[ , -1]

#create heatmap with pheatmap
pheatmap(heatmap_data,
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         display_numbers = TRUE,
         number_format = "%.0f",
         fontsize_number = 6,
         fontsize_row = 10,
         fontsize_col = 10,
         main = "Gene Expression Heatmap by Age",)
```



Part 4: A plot type we did not discuss in class

```
#percent of people in the ICU on mechanical ventilation

df$mechanical_ventilation <- trimws(df$mechanical_ventilation)

on_vent <- df %>%
  group_by(mechanical_ventilation) %>%
  summarize(count = n())

print(on_vent)
```

```
## # A tibble: 2 x 2
##   mechanical_ventilation count
##   <chr>                  <int>
## 1 no                        12
## 2 yes                       29
```

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

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

```
pie_chart <- ggplot(on_vent, aes(x = "", y = count, fill = mechanical_ventilation)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
```

```
  labs(title = "Mechanical Ventilation Count Among ICU Patients") +
  theme_void() +
  scale_fill_manual(values = c("yes" = "darkorange1", "no" = "cornflowerblue")) +
   geom_text(aes(label = count),
             position = position_stack(vjust = 0.5))

print(pie_chart)
```