

Final Project Presentation 1

Sai Lakkireddy

2023-07-24

Chosen Gene: AAMP

AAMP stands for “Angio-Associated Migratory Cell Protein” gene. This gene is responsible for making the AAMP protein, which is involved in cell movement and angiogenesis, contributing to processes like wound healing and tissue repair. Genes and their corresponding proteins are crucial for the proper functioning of our bodies.

In this analysis, we study the association between AAMP’s gene expression, age, COVID Status and ICU Status

Importing and combining the Data from two csv files

Steps:

1. Import both the csv files
2. Convert the gene expression into long format
3. Inner join it with meta data

```
#set current working directory to the previous folder
setwd("../")

#import the geneExpression and metaData csv from the data folder
data_gene_Expression <- read.csv("data/QBS103_finalProject_geneExpression.csv", header=TRUE)
data_meta <- read.csv("data/QBS103_finalProject_metadata.csv", header=TRUE)

#We convert the geneExpression data from wide form into long form
data_gene_Expression.longForm <- data_gene_Expression %>%
  pivot_longer(cols = starts_with(c("COVID_", "NONCOVID_")),
    names_to = "participant_id",
    values_to = "gene_expression_value"
  )

#make a final data frame by combining the two data sets and making it a data frame
final_df <- as.data.frame(data_gene_Expression.longForm %>% inner_join( data_meta,
  by=c('participant_id')))
```

Pre-processing the data

Steps:

1. Remove “unknown” strings and prefixes
2. Convert the class the columns to their appropriate type

```
#rename with x column with gene
final_df <- rename(final_df, gene = X)

#remove all unknown strings and substitute it with NAs
final_df[, 16:27][final_df[, 16:27] == ' unknown' | final_df[, 16:27] == 'unknown'] <- NA

#format the disease status column to just include the status
final_df$disease_status <- sub('disease state: ', '', final_df$disease_status)

#convert the column type of disease_status, sex, icu_status and mechanical_ventilation to factor
final_df <- final_df %>%
  mutate_at(vars(disease_status, sex, icu_status, mechanical_ventilation), as.factor)

#convert the class of age, charlson_score
final_df <- final_df %>%
  mutate_at(vars(age, apacheii, ferritin.ng.ml.,
                 crp.mg.l., ddimer.mg.l_feu.,
                 procalcitonin.ng.ml., lactate.mmol.l., fibrinogen, sofa), as.integer)
```

Optional - handle missing values

```
# check all the numeric columns
num_cols <- names(select_if(final_df, is.numeric))

# Create an imputation model
imputation_model <- mice(final_df[num_cols], method = "pmm", printFlag = FALSE)

# Perform the imputation
imputed_data_final <- complete(imputation_model)
# update the final data frame with the imputed values
final_df[num_cols] <- imputed_data_final[num_cols]
```

Create a subset the AAMP Gene and the chosen covariates

```
final_subset <- final_df[final_df$gene == 'AAMP',
  c('gene',
    'gene_expression_value', 'age',
    'icu_status', 'disease_status')]

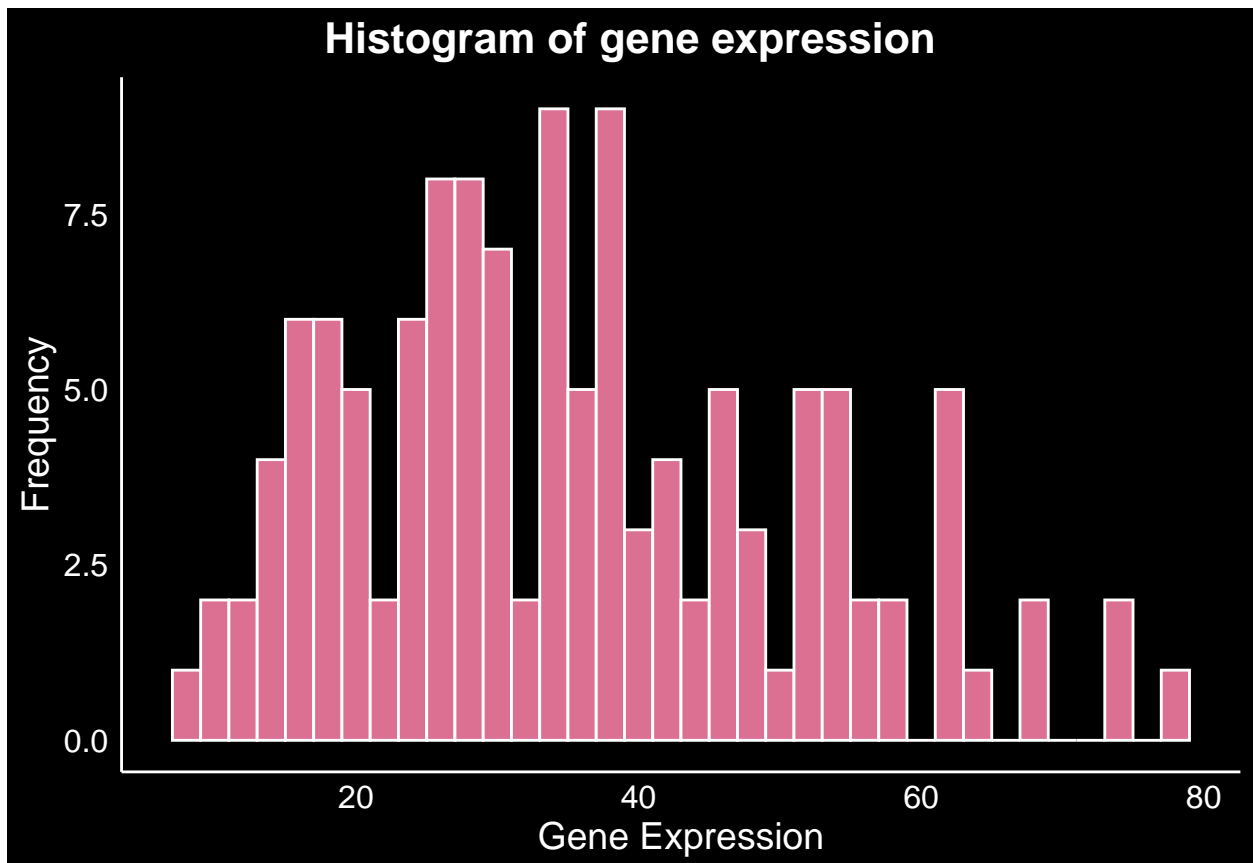
head(final_subset)
```

```
##      gene gene_expression_value age icu_status disease_status
```

| | | | | | | |
|----|------|------|-------|----|----|----------|
| ## | 4501 | AAMP | 61.08 | 39 | no | COVID-19 |
| ## | 4502 | AAMP | 54.54 | 63 | no | COVID-19 |
| ## | 4503 | AAMP | 25.19 | 33 | no | COVID-19 |
| ## | 4504 | AAMP | 67.95 | 49 | no | COVID-19 |
| ## | 4505 | AAMP | 18.29 | 49 | no | COVID-19 |
| ## | 4506 | AAMP | 51.35 | 38 | no | COVID-19 |

Histogram for Gene Expression

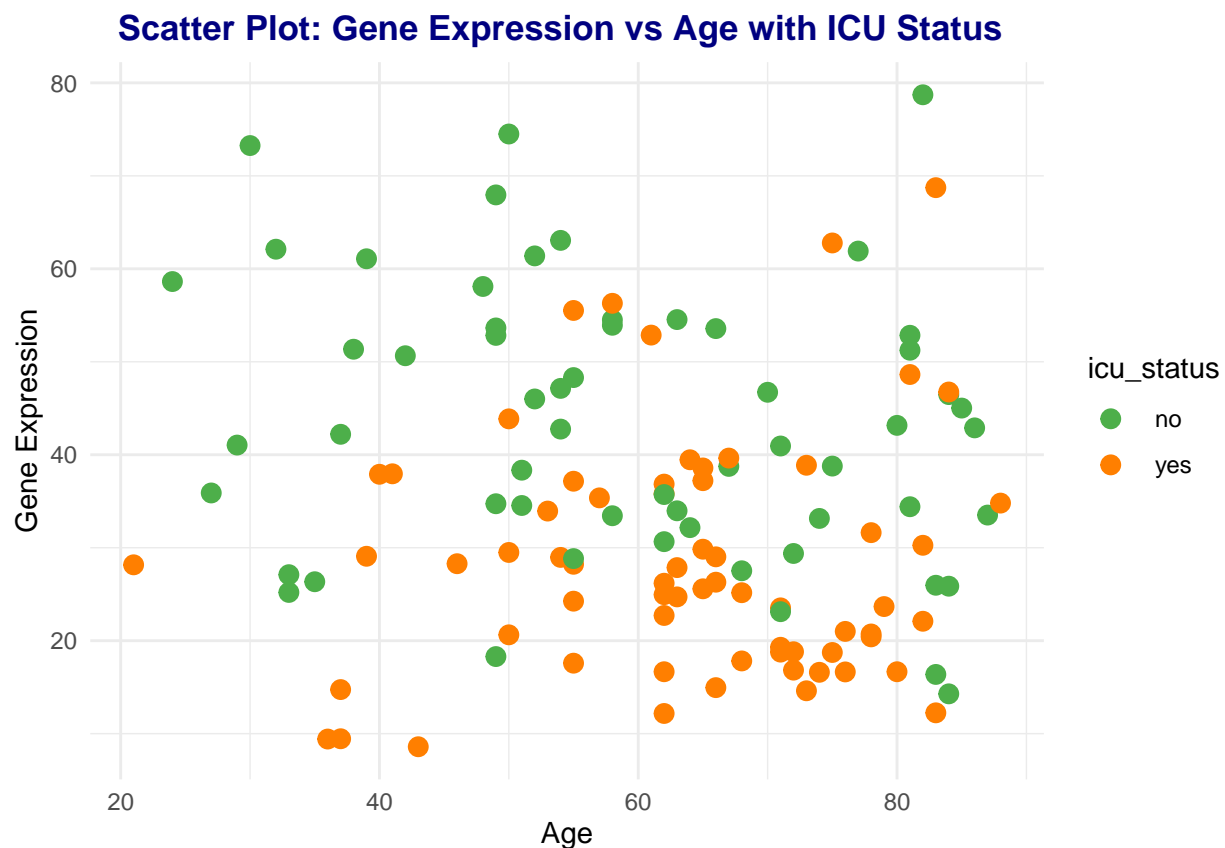
```
ggplot(final_subset, aes(x = gene_expression_value)) +
  geom_histogram(binwidth = 2, color = "white", fill = "#DB7093") +
  theme_minimal() +
  theme(
    panel.background = element_rect(fill = "black"),
    plot.background = element_rect(fill = "black"),
    axis.line = element_line(color = "white"),
    axis.text = element_text(color = "white", size = 12),
    axis.title = element_text(color = "white", size = 14),
    panel.grid = element_blank(),
    plot.title = element_text(color = "white", size = 16, face = "bold", hjust = 0.4),
  ) +
  ggtitle("Histogram of gene expression") +
  xlab("Gene Expression") +
  ylab("Frequency")
```



Scatter plot: Age vs Gene Expression factoring for ICU status

```
my_colors_1 <- c("#4DAF4A", "#FF7F00")

# Create the scatter plot with custom color scheme
ggplot(final_subset, aes(x = age, y = gene_expression_value, color = icu_status)) +
  geom_point(size = 3) +
  scale_color_manual(values = my_colors_1) +
  theme_minimal() +
  theme(
    plot.title = element_text(color = "navy", size = 13, face = "bold", hjust = 0.4)
  ) +
  ggtitle("Scatter Plot: Gene Expression vs Age with ICU Status") +
  xlab("Age") +
  ylab("Gene Expression")
```



Box plot: Gene Expression by COVID and ICU Status

```
my_colors_2 <- c("#377eb8", "#e41a1c")
ggplot(final_subset, aes(x = disease_status, y = gene_expression_value, fill = icu_status)) +
  geom_boxplot(color = "black", width = 0.5, alpha = 0.8) +
```

```

scale_fill_manual(values = my_colors_2) +
theme_minimal() +
  theme_minimal() +
    theme(
      plot.title = element_text(color = "darkgreen", size = 13, face = "bold", hjust = 0.4)
    )+
ggtitle("Box plot of Gene Expression by COVID and ICU Status") +
xlab("Disease Status") +
ylab("Gene Expression")

```

