My_Graphs

2024-07-25

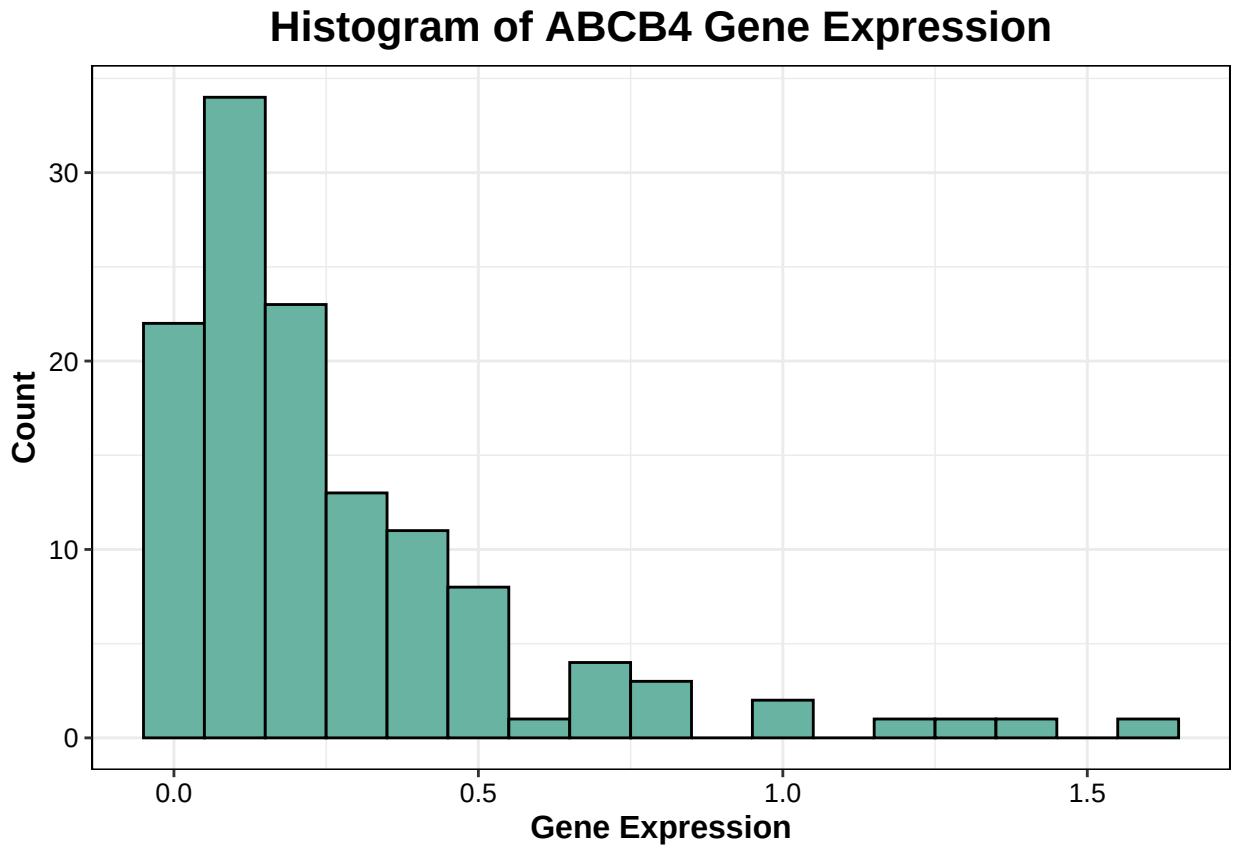
```
# Load packages and tidyverse includes ggplot2
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

```
# Pulling Data from the two provided files.
# row.names = 1 to indicate that the first column of the CSV file is names for the data frame.
gene_express_df <- read.csv("C:/Users/casha/Downloads/QBS103_GSE157103_genes.csv", row.names = 1)
metadata_df <- read.csv("C:/Users/casha/Downloads/QBS103_GSE157103_series_matrix.csv")

# Selecting specific columns of interest
metadata_relevant <- metadata_df %>%
  select(participant_id, geo_accession, age, disease_status, sex)
#Continuous Covariant = age
#Categorical Covariants = disease_status, sex

# Extract expression data for the ABCB4 gene
abcb4_expression <- gene_express_df %>%
  rownames_to_column("gene") %>%
  filter(gene == "ABCB4") %>% #Keep ONLY ABCB4 rows

#wide format to long format
  pivot_longer(cols = -gene, names_to = "participant_id", values_to = "expression") %>%
  select(-gene)

# Merge expression data with metadata
merged_df <- abcb4_expression %>%
#https://www.datacamp.com/tutorial/merging-data-r: How to merge 2 datasets in R
  merge(metadata_relevant, by = "participant_id")

# Ensure correct data types
merged_df <- merged_df %>%
  mutate(age = as.numeric(age), #make sure age is numeric
         sex = factor(sex),
         disease_status = factor(disease_status))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `age = as.numeric(age)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# Histogram of ABCB4 gene expression
histogram <- ggplot(merged_df, aes(x = expression)) +
#https://www.datacamp.com/tutorial/make-histogram-basic-r
  geom_histogram(binwidth = 0.1, fill = "#69b3a2", color = "black") + #binwidth is how wide we want eac
#changed colors to make it more like other plots I've seen in papers

#change labels
  labs(
    title = "Histogram of ABCB4 Gene Expression",
    x = "Gene Expression",
    y = "Count" #number which falls within THIS level of gene expression
  ) +
  theme_bw() +
  theme(
    plot.background = element_rect(fill = "white"),
    panel.background = element_rect(fill = "white"),
    panel.border = element_rect(color = "black", fill = NA),
    text = element_text(size = 12),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(color = "black", size = 10)
  )

histogram
```




```
# Scatterplot of ABCB4 expression vs. age
scatterplot <- ggplot(merged_df, aes(x = age, y = expression)) +
  geom_point(color = "#1F77B4", size = 3, alpha = 0.7) +  #using a clear color and changing the thickne
  geom_smooth(method = "loess", color = "#69b3a2", se = FALSE, size = 1) + #keeping the color scheme si
  # Add a regression line https://stackoverflow.com/questions/15633714/adding-a-regression-line-on-a-gg
# (decided to go with a loess instead of lm since a loess line shows the higher average gene expression
  labs(
    title = "Scatterplot of ABCB4 Gene Expression vs Age",
    x = "Age (yrs)", #continuous covariate
    y = "Gene Expression"
  ) +
  theme_bw() + #same theme for all graphs
  theme(
    plot.background = element_rect(fill = "white"),
    panel.background = element_rect(fill = "white"),
    text = element_text(size = 12),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16), #keeping sizing consistent betwee
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(color = "black", size = 10),
    panel.border = element_rect(color = "black", fill = NA)
#https://www.geeksforgeeks.org/add-panel-border-to-ggplot2-plot-in-r/
#don't fill border or no graph appears
#    panel.grid = element_line(color = "grey") : decided to remove as did not help but learned a new te
#https://r-charts.com/ggplot2/grid/
  )
```

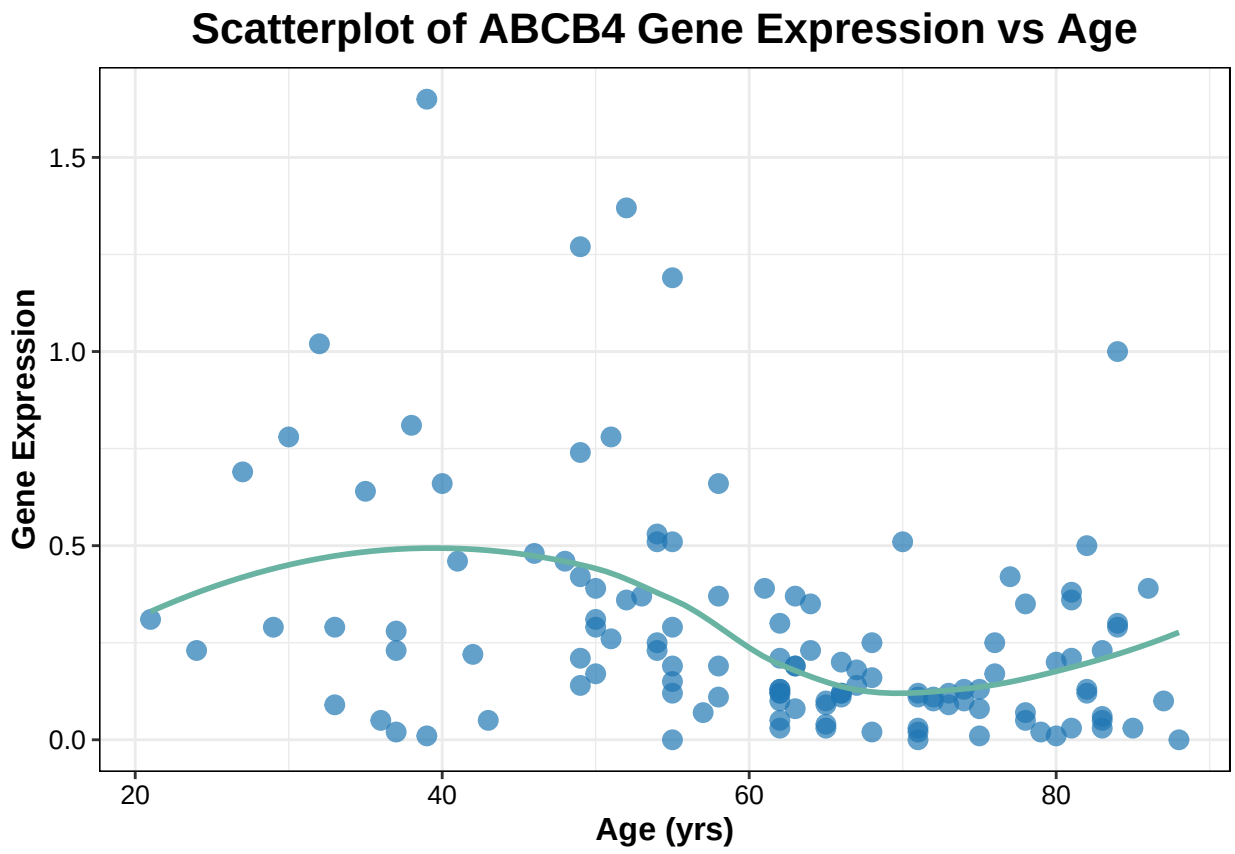
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
scatterplot
```

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

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (‘stat_smooth()‘).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



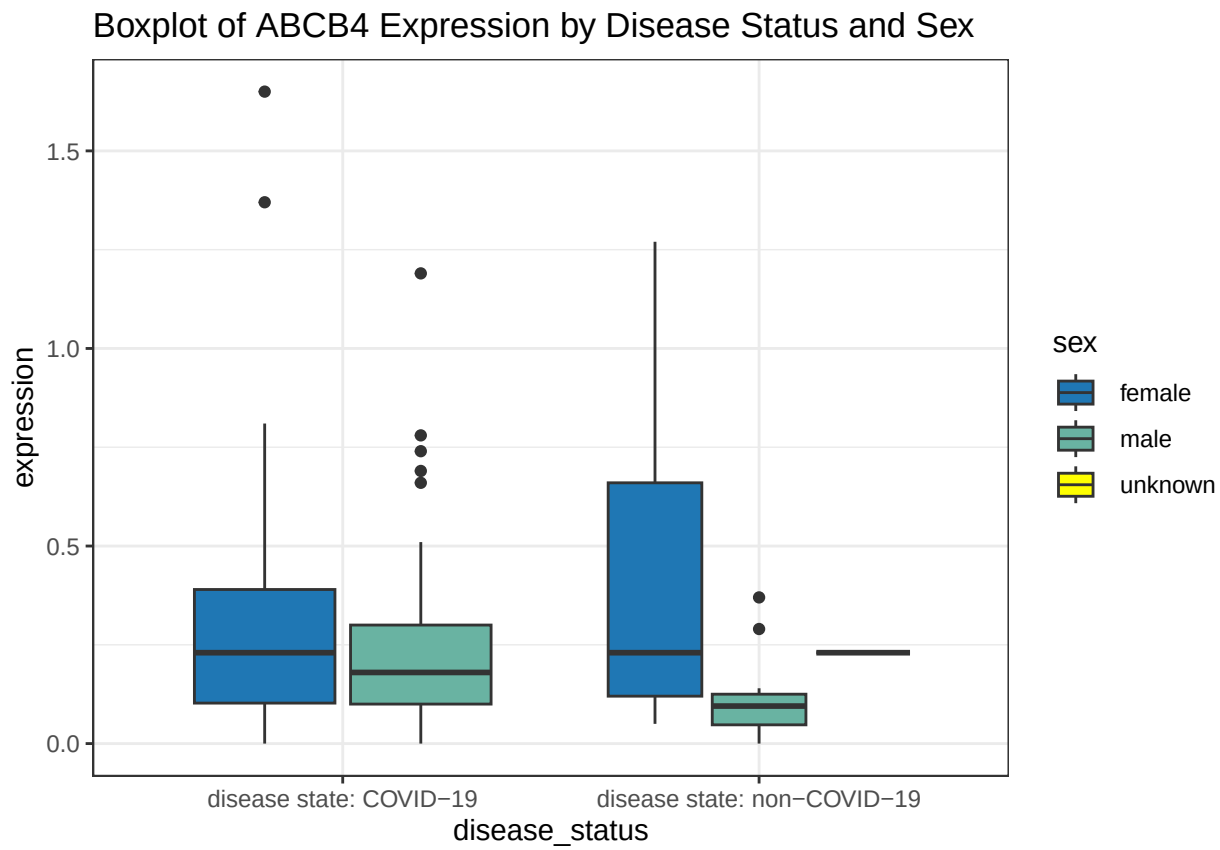

```
#Boxplot
boxplot <- ggplot(merged_df,aes(x = disease_status, y = expression, fill = sex)) +
#Add box plot
geom_boxplot() +
#Define colors: Have to use three colors because sex is female, male, and unknown
scale_fill_manual(values = c('#1F77B4', '#69b3a2', 'yellow')) +
#using the same color scheme
    labs(
```

```
    title = "Boxplot of ABCB4 Expression by Disease Status and Sex"
  ) +
  theme_bw() #same theme
  theme(
    text = element_text(size = 12),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16), #keeping sizing consistent betwee
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(color = "black", size = 10),
    panel.border = element_rect(color = "black", fill = NA)
  )
```

```
## List of 5
##  $ text        :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : num 12
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.title  :List of 11
##   ..$ family       : NULL
##   ..$ face         : chr "bold"
##   ..$ colour       : NULL
##   ..$ size         : num 12
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ axis.text   :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : chr "black"
##   ..$ size         : num 10
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ panel.border:List of 5
##   ..$ fill         : logi NA
```

```
##   ..$ colour       : chr "black"
##   ..$ linewidth    : NULL
##   ..$ linetype     : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_rect" "element"
##  $ plot.title  :List of 11
##   ..$ family       : NULL
##   ..$ face         : chr "bold"
##   ..$ colour       : NULL
##   ..$ size         : num 16
##   ..$ hjust        : num 0.5
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

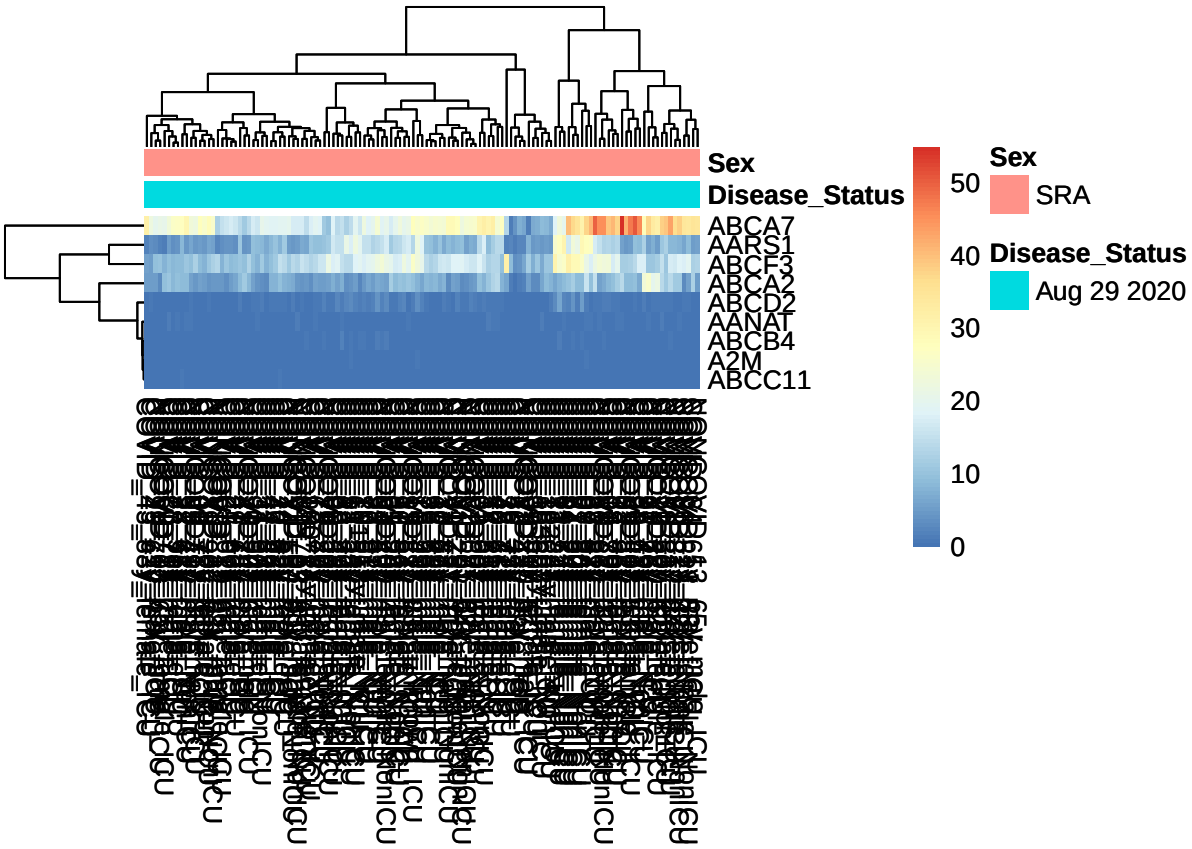
```
boxplot
```



```r
library(pheatmap)
# MANY ways to make a heatmap. Googled anything I was unsure about and linked the results in the docume

#https://davetang.org/muse/2018/05/15/making-a-heatmap-in-r-with-the-pheatmap-package/ - How to make a

# Calling the data under different names for the sake of simplicity
genes_df <- read.csv("C:/Users/casha/Downloads/QBS103_GSE157103_genes.csv", row.names = 1)
metadata_df <- read.csv("C:/Users/casha/Downloads/QBS103_GSE157103_series_matrix.csv", sep = "\t", head

# Pick 10 genes of interest
genes_of_interest <- c("A2M", "ABCB4", "AANAT", "AARS1", "ABCA2", "ABCB4", "ABCC11", "ABCD2", "ABCA7",
filtered_genes_df <- genes_df[rownames(genes_df) %in% genes_of_interest, ]

# Separate the metadata
metadata_processed <- metadata_df %>%
  separate(V1, into = c("Sample", "GSM", "Public", "Source_Name", "Disease_Status", "Sex", "Additional_
  select(Sample, Disease_Status, Sex)
```

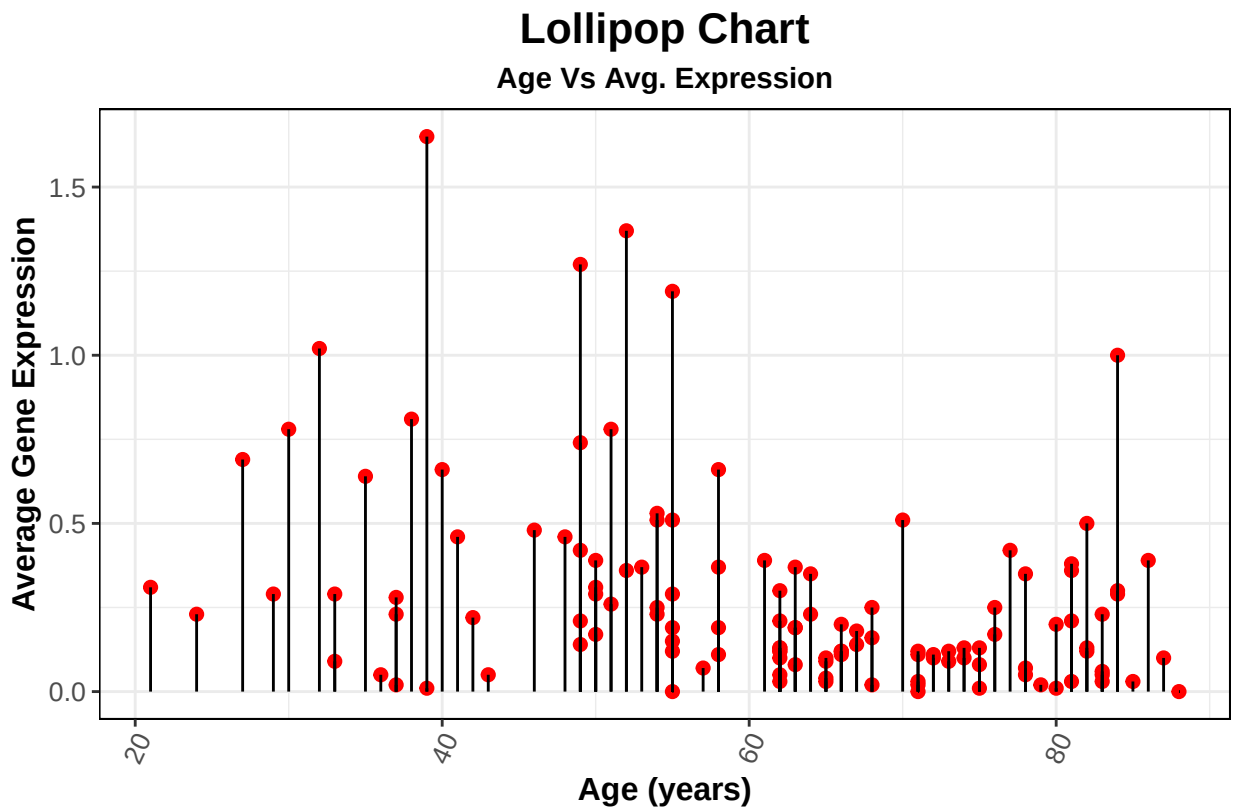
```
## Warning: Expected 7 pieces. Additional pieces discarded in 127 rows [1, 2, 3, 4, 5, 6,
## 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```r
# Remove rows with NA values in essential columns or will cause errors
metadata_processed <- metadata_processed %>% drop_na(Sample, Disease_Status, Sex)
# Check and match the number of columns
metadata_processed <- metadata_processed %>%
  filter(Sample %in% colnames(filtered_genes_df))

# Ensure that the sample names in metadata match the column names in the gene expression data
filtered_genes_df <- filtered_genes_df[, metadata_processed$Sample]

# Create annotation data
annotation <- data.frame(
  Disease_Status = metadata_processed$Disease_Status,
  Sex = metadata_processed$Sex
)
rownames(annotation) <- metadata_processed$Sample

# Generate the heatmap using pheatmap
heatmap <- pheatmap(

  as.matrix(filtered_genes_df),
  annotation_col = annotation,
# How to cluster rows and columns in a heatmap https://www.geeksforgeeks.org/draw-heatmap-with-clusters
  clustering_distance_rows = "euclidean",
  clustering_distance_cols = "euclidean",
  show_rownames = TRUE,
  show_colnames = TRUE
)


heatmap
```




```
# List of possible ggplot2 plots? - https://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-C
theme_set(theme_bw())

# Plot
lollipop <- ggplot(merged_df, aes(x = age, y = expression)) +
  geom_point(size=2, color = "red") + #most resembles a lollipop color
  geom_segment(aes(x=age,
                   xend=age,
                   y=0,
                   yend=expression)) +
  labs(title="Lollipop Chart",
       x = "Age (years)",
       y = "Average Gene Expression",
       subtitle="Age Vs Avg. Expression",
       caption="source:QBS103_GSE157103") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6),
    plot.background = element_rect(fill = "white"),
    panel.background = element_rect(fill = "white"),
    panel.border = element_rect(color = "black", fill = NA),
    text = element_text(size = 12),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    plot.subtitle = element_text(hjust = 0.5, face = "bold", size = 11),
    axis.title = element_text(face = "bold", size = 12),
  )

lollipop
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).

## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_segment()`).
```



source:QBS103_GSE157103

```
# Install and load packages
# install.packages("knitr")
# install.packages("kableExtra")
# install.packages("dplyr")
```

```
# Load necessary libraries
# https://bookdown.org/yihui/rmarkdown-cookbook/kable.html
library(dplyr)
library(knitr)
library(kableExtra)  # Ensure kableExtra is loaded for kable_styling()
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```r
# Load necessary libraries
library(dplyr)
library(knitr)

# Load your dataset (assuming you have the file in the working directory)
series_matrix_df <- read.csv("C:/Users/casha/Downloads/QBS103_GSE157103_series_matrix.csv", stringsAsFa

#colnames(series_matrix_df) checking names because of errors

# Convert relevant columns to numeric where needed
series_matrix_df$age <- as.numeric(gsub("[^0-9]", "", series_matrix_df$age))
series_matrix_df$ferritin.ng.ml <- as.numeric(series_matrix_df$`ferritin.ng.ml.`)
```

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

```r
series_matrix_df$ddimer.mg.l_feu <- as.numeric(series_matrix_df$`ddimer.mg.l_feu.`)
```

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

```r
# View summary to check for any issues
# summary(series_matrix_df)

# Group by sex and calculate summary statistics
summary_stats <- series_matrix_df %>%
  group_by(sex) %>%
  summarise(
    age_mean = mean(age, na.rm = TRUE),
    age_sd = sd(age, na.rm = TRUE),
    ferritin_median = median(ferritin.ng.ml, na.rm = TRUE),
    ferritin_iqr = IQR(ferritin.ng.ml, na.rm = TRUE),
    ddimer_median = median(ddimer.mg.l_feu, na.rm = TRUE),
    ddimer_iqr = IQR(ddimer.mg.l_feu, na.rm = TRUE),
    disease_status_n = n(),
    mechanical_ventilation_n = n()
  )

# Add percentages for categorical variables
summary_stats <- summary_stats %>%
  mutate(
    disease_status_percent = (disease_status_n / sum(disease_status_n)) * 100,
    mechanical_ventilation_percent = (mechanical_ventilation_n / sum(mechanical_ventilation_n)) * 100
  )

# View the summary statistics
# print(summary_stats)

# Render the summary statistics table using kable
# How to construct a nice summary table in r? - https://rdrr.io/github/grayclhn/dbframe-R-library/man/b
kable(summary_stats, format = "latex", booktabs = TRUE, caption = "Summary Statistics Stratified by Sex
  kable_styling(latex_options = c("striped", "hold_position"))
```

Table 1: Summary Statistics Stratified by Sex

sex	age_mean	age_sd	ferritin_median	ferritin_iqr	ddimer_median	ddimer_iqr	disease_status_n	r
female	59.88235	18.22158	318	547	1.37	5.86	51	
male	62.64384	14.64901	755	849	2.21	10.55	74	
unknown	83.00000	NA	NA	NA	NA	NA	1	