amy_wang_QBS_project

2024-07-25

```
#importing the csv files
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
```

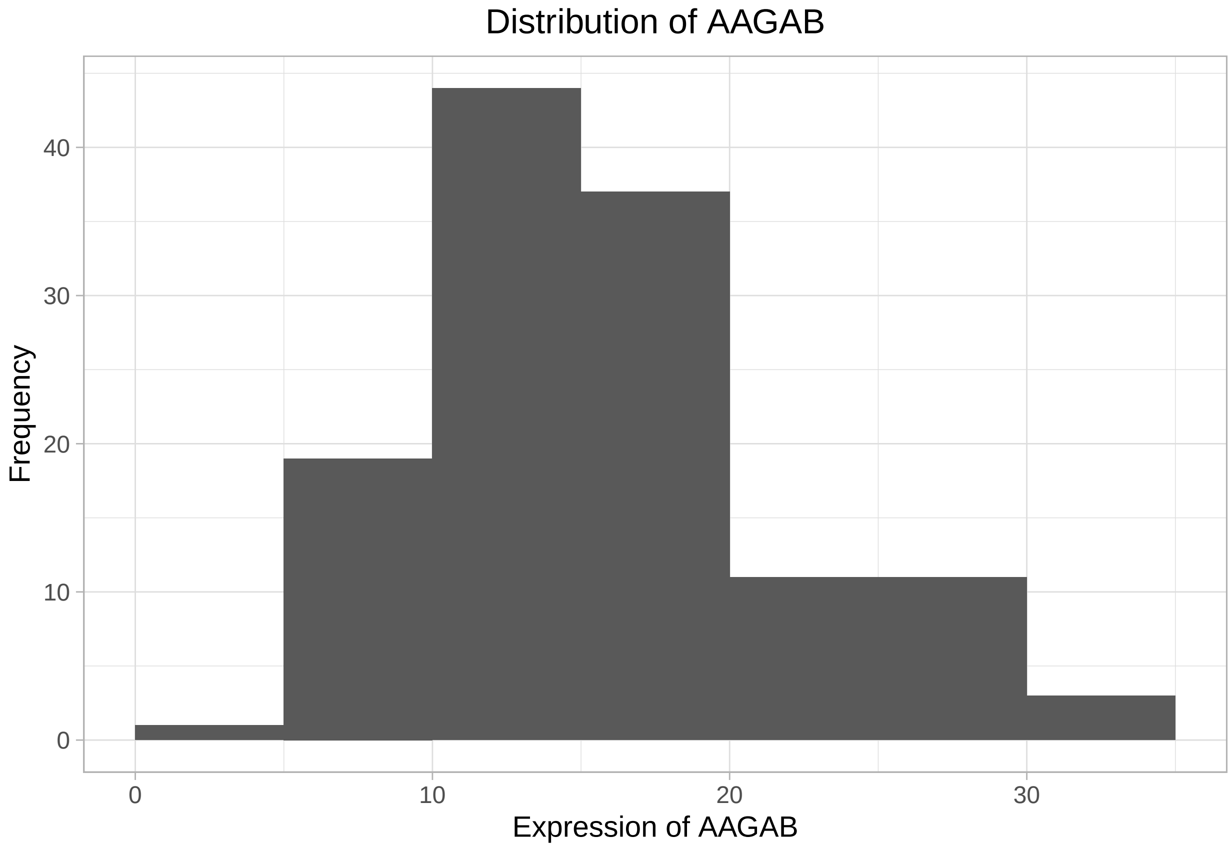
```
datafiles <- list.files(path = "/Users/amyw/Documents/QBS103/DSProject", pattern = ".csv")
print(datafiles)
```

```
## [1] "QBS103_GSE157103_genes.csv"         "QBS103_GSE157103_series_matrix.csv"
```

```
setwd("/Users/amyw/Documents/QBS103/DSProject")

for (file in datafiles){
  the_file <- read.csv(file)
  #print(head(the_file)) #commented out for ease of marking
}

#gene: AAGAB
#continuous variate: age
#categorical covariates: sex, mechanical ventilation
```

```
#creating new dataframes

setwd("/Users/amyw/Documents/QBS103/DSProject")

gene_df <- read.csv('QBS103_GSE157103_genes.csv', header = TRUE,row.names = 1)
meta_df <- read.csv('QBS103_GSE157103_series_matrix.csv', header = TRUE, row.names = 1)

#print(head(gene_df)) #for ease of marking
#print(head(meta_df))
```

```
#plotting histogram of AAGAB
working_genes <- as.data.frame(t(gene_df)) #pivoting dataframe to select AAGAB as a column
#print(head(working_genes))

ggplot(working_genes, aes(x=AAGAB))+
  geom_histogram(binwidth = 5, boundary = 0)+ #boundary at 0 to group data
  labs(title = 'Distribution of AAGAB', x = 'Expression of AAGAB', y = 'Frequency')+
  theme_light()+
  theme(plot.title = element_text(hjust=0.5)) #centering title
```



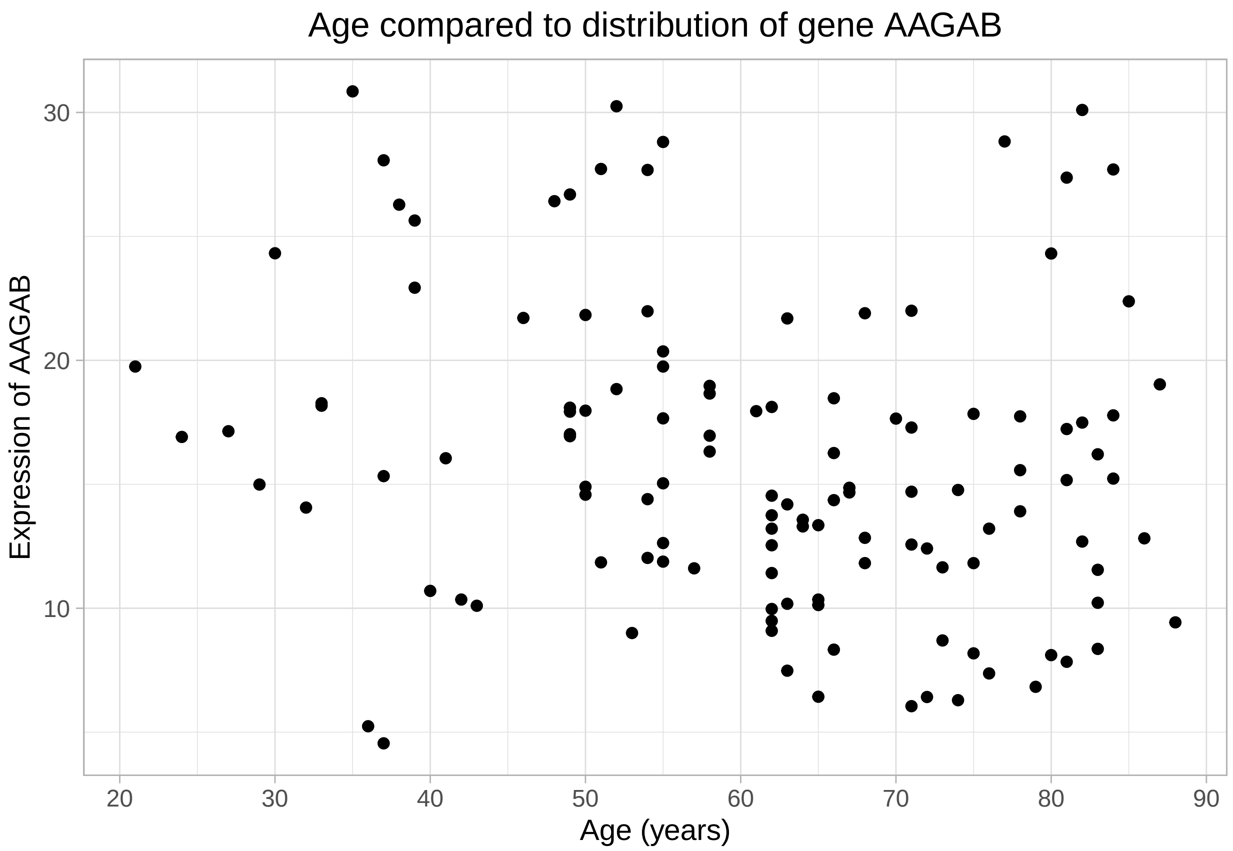
```
#merging the data into new dataframe called 'new_comp' and reformatting to 'final_comp'

#print(head(working_genes))
#print(head(meta_df))

new_comp <- merge(working_genes, meta_df, by = 'row.names') #merging the two dataframes using the row n
final_comp <- data.frame(new_comp, row.names = 1) #renaming the row names as the participant ID
#print(head(final_comp))
```

```
#plotting scatter plot

final_comp$age<- as.numeric(final_comp$age) #converting the age column from characters to numerics
```

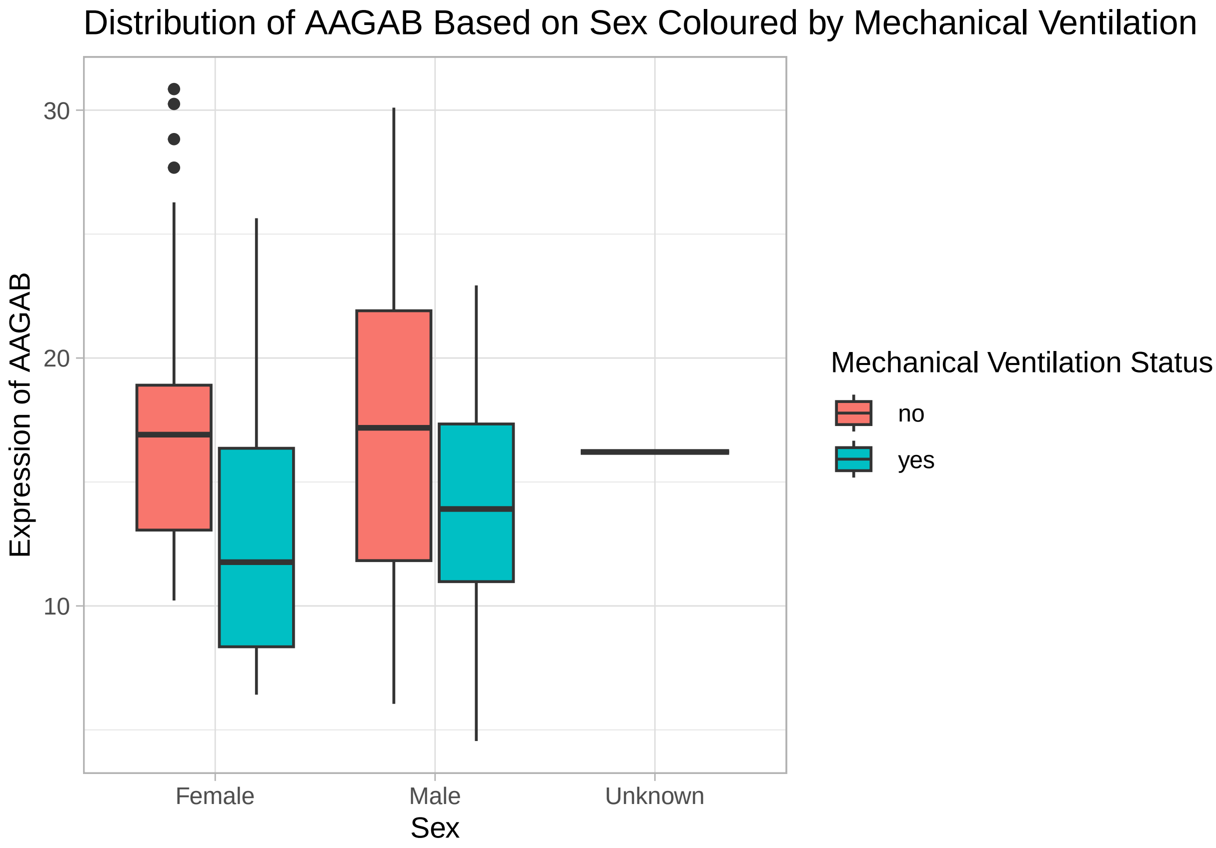
```
## Warning: NAs introduced by coercion
```

```
ggplot(final_comp, aes(x= age, y=AAGAB))+
  geom_point()+ #two age unknowns from the data
  labs(title = 'Age compared to distribution of gene AAGAB',
       x = 'Age (years)', y = 'Expression of AAGAB')+
  scale_x_continuous(breaks=seq(0,150,by=10))+
  theme_light()+
  theme(plot.title = element_text(hjust=0.5))
```

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



```
#print(head(final_comp))
box_first <- dplyr::select(final_comp,AAGAB, sex, mechanical_ventilation) #selecting data columns for e
#print(box_first)

box_long <- box_first %>% #pivoting long to prep for boxplot
  tidyr::pivot_longer(cols= c(sex, mechanical_ventilation),
                      names_to = 'categorical', values_to = 'Sex/MechVent')
#head(box_long)

ggplot(box_first, aes (sex, AAGAB, fill = mechanical_ventilation))+
  geom_boxplot()+
  theme_light()+
  labs(title = 'Distribution of AAGAB Based on Sex Coloured by Mechanical Ventilation',
       x = 'Sex', y = 'Expression of AAGAB')+
  theme(plot.title = element_text(hjust=0))+
  scale_fill_discrete(name = 'Mechanical Ventilation Status')+
  scale_x_discrete(labels = c('Female', 'Male', 'Unknown'))
```



```
final_df <- as.data.frame(t(final_comp))
#print(head(final_df))
#reformatting dataframe back to industry standard of having participant ID as the column names
```