Final Project Submission 1

Vui Doan

2025-07-14

```
################################################
# Reading gene expression data with read.csv2. #
################################################

# The field separator is set to comma.

# The argument row.names is set to 1 so that when the file is read the firt column as row.names of the

genes_data <- read.csv2(file = "C:/Users/Jade/Desktop/GitRepo/Repository_One/Gene_expression_QBS103_GSE

# Cheking dimension data.frame

dim(genes_data)
```

```
## [1] 100 126
```

```
#generates visualization of the data frame similar to an excel table

# View(genes_data)

#used to see the names of each row

# rownames(genes_data)

# colnames(genes_data)

# rownames(genes_data)

# colnames(genes_data)

#Shows the whole row

# genes_data["A1BG",]

# genes_data[c("ABCG2","ABHD1"),]

#shows whole column

# genes_data[,"COVID_01_39y_male_NonICU"]
```

```r
# read.scv2 will load the data into a data.frame
covariates_data <- read.csv2(file = "C:/Users/Jade/Desktop/GitRepo/Repository_One/Covariates_QBS103_GSE
#view covariates in columns and select one
# colnames(covariates_data)
# to access specific rows and columns. First number in bracket is row and second bracket is column.
# covariates_data[1,]
# covariates_data[,2]
#to select multiple rows create a vector with all of the rows you want to select
# covariates_data[c(1,3,5),]
# covariates_data[,c(5,7)]
# covariates_data[c(1,2),c(6,7)]
# View(covariates_data)
# covariates_data[,"disease_status"]
#continuous:
#categorical: Sex and disease status
slected_covariates <- c("hospital.free_days_post_45_day_followup","disease_status","sex")
covariates_data_filt <- covariates_data[,slected_covariates]
# Selecting the first gene.
# data frame with one row
selected_gene <- genes_data[1,]
#combine expression of gene selected and the information of the covariates
#viewed/printed the number or rows and columns (dimensions) of the covariates selected
dim(covariates_data_filt)
```

```
## [1] 126   3
```

```r
#the genes were in rows and covariates in columns, so selected gene's row was transposed to match orien
selected_gene_tran <- t(selected_gene)
#look at the dimensions to make sure it matches
```

```r
dim(selected_gene_tran)
```

```
## [1] 126   1
```

```r
data <- cbind(covariates_data_filt,selected_gene_tran)

#cbind =column bind.

# help(cbind)

# View(data)

class(data[,1])
```

```
## [1] "integer"
```

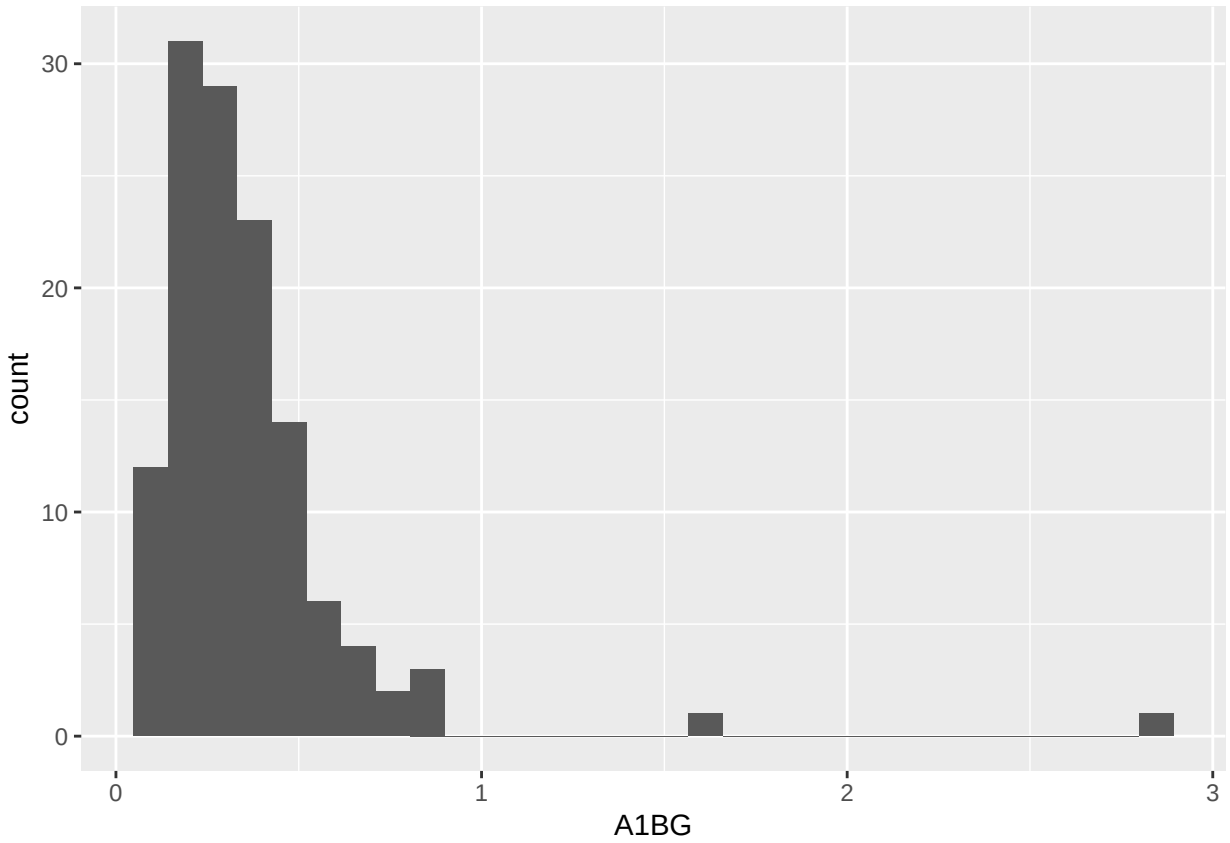
```r
data[,4] <- as.numeric(data[,4])

class(data[,4])
```

```
## [1] "numeric"
```

```r
####################
#CREATING HISTOGRAM#
####################

# data[,"A1BG"]

# tinytex::install_tinytex()
```

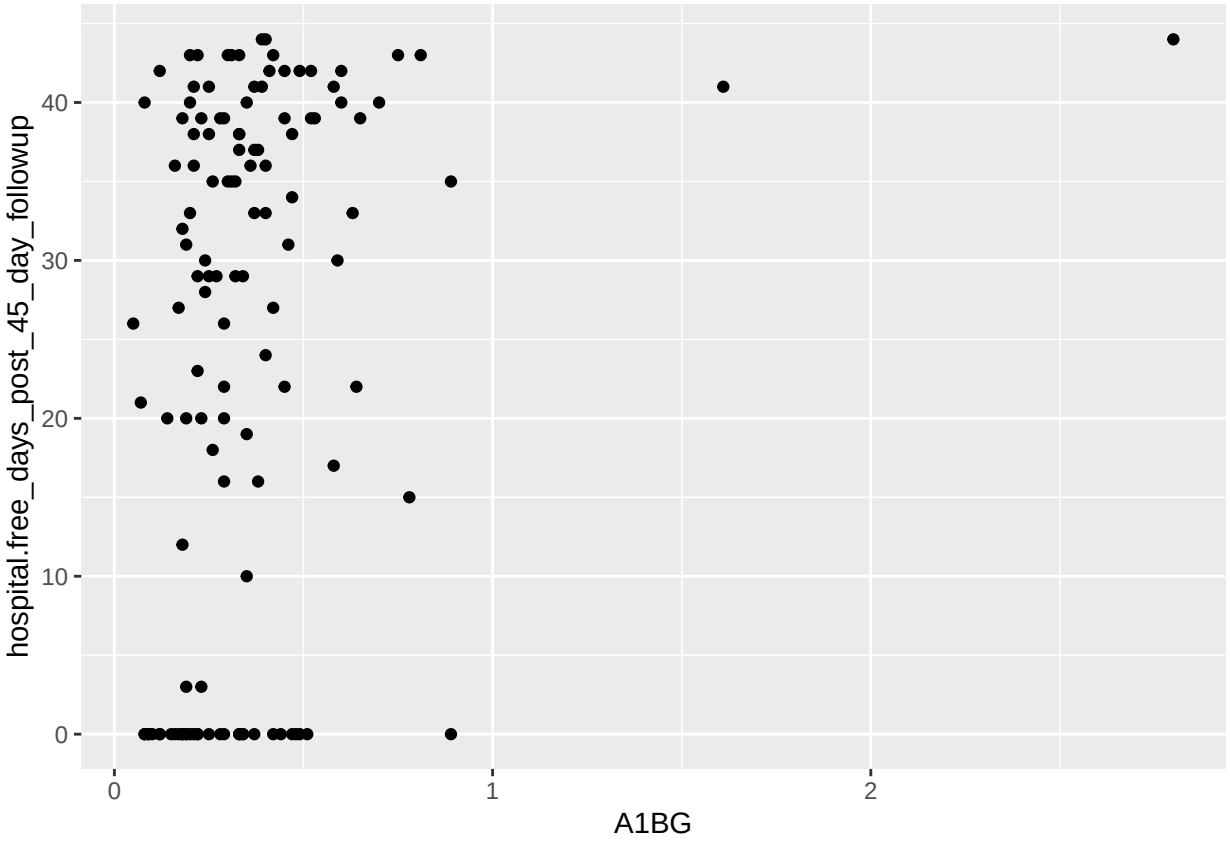
```r
# install.packages("ggplot2")

library(ggplot2)

ggplot(data, aes(x = A1BG)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

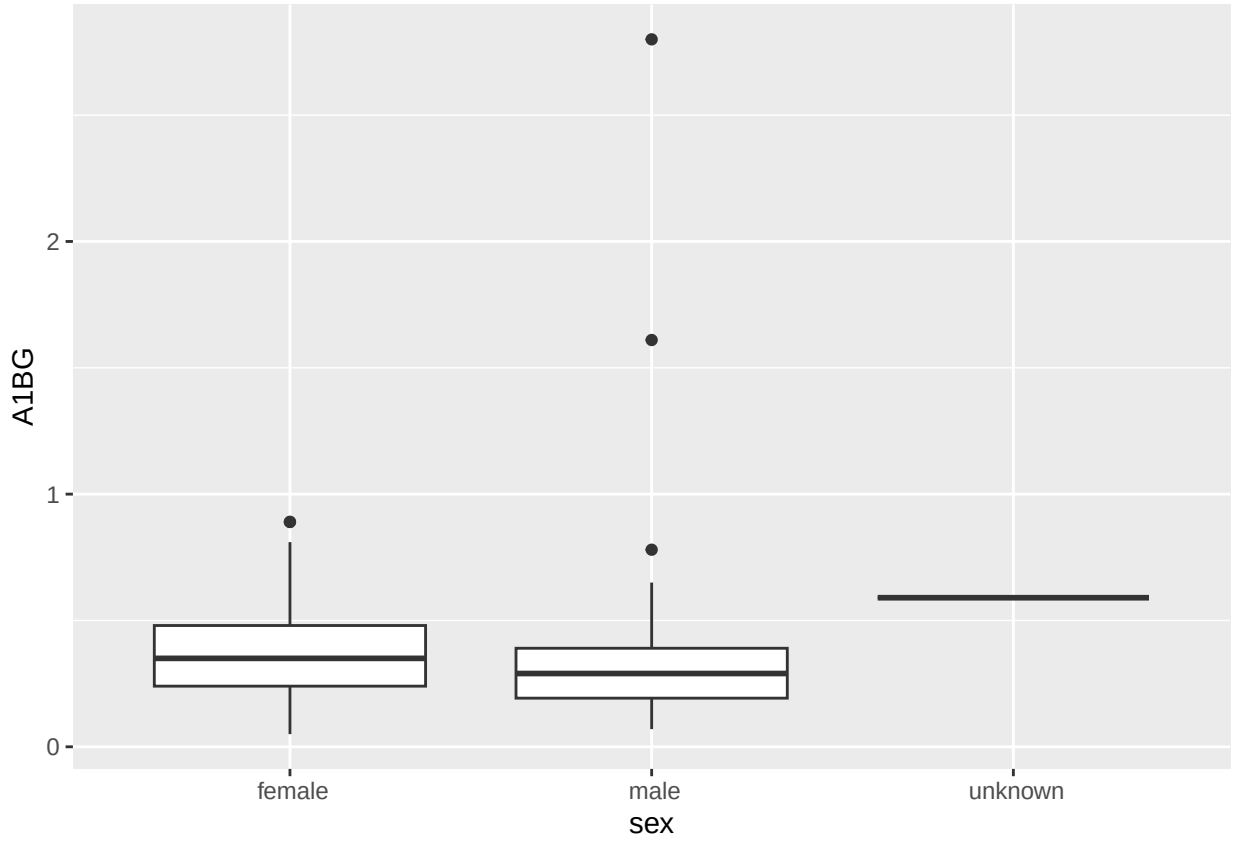



```
#The histogram shows the number of samples that present expression levels of the gene inside a particul
```

```
ggplot(data, aes(x = A1BG, y = hospital.free_days_post_45_day_followup)) + geom_point()
```

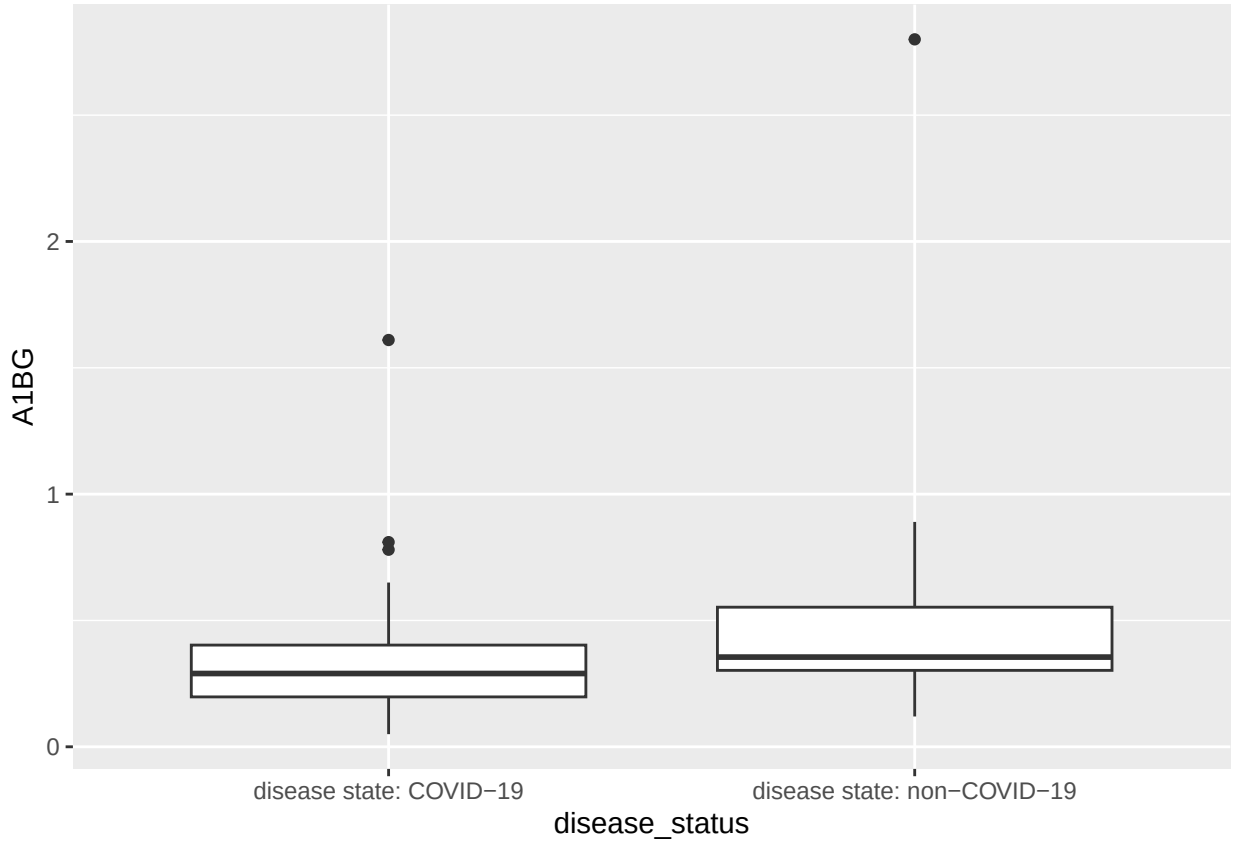



```
ggplot(data, aes(x = sex , y = A1BG)) + geom_boxplot()
```



FOR THE UNKNOWN, ATLEAST ONE OF THE SEX WAS NOT REPORTED

```
# colnames(data)

# tinytex::install_tinytex(force = TRUE)

ggplot(data, aes(x = disease_status , y = A1BG)) + geom_boxplot()
```


A1BG
2
1
0
disease state: COVID−19
disease state: non−COVID−19
disease_status