Ruiqi_Li_Submission1

Ruiqi Li

2025-07-13

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
#Load gene expression data
gene_data <- read.csv(file = "QBS103_GSE157103_genes.csv",row.names=1)
dim(gene_data)
```

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

```r
str(gene_data)
```

```
## ’data.frame':    100 obs. of  126 variables:
##  $ COVID_01_39y_male_NonICU     : num  0.49 0 0.21 0.04 0.07 ...
##  $ COVID_02_63y_male_NonICU     : num  0.29 0 0.14 0 0 ...
##  $ COVID_03_33y_male_NonICU     : num  0.26 0 0.03 0.02 0 ...
##  $ COVID_04_49y_male_NonICU     : num  0.45 0.01 0.09 0.07 0 ...
##  $ COVID_05_49y_male_NonICU     : num  0.17 0 0 0.05 0.07 0 0 8.45 1.17 0 ...
##  $ COVID_06_.y_male_NonICU      : num  0.21 0 0.08 0.04 0 0 0.03 19.6 3.15 0 ...
##  $ COVID_07_38y_female_NonICU   : num  0.49 0.01 0.23 0.03 0.07 ...
##  $ COVID_08_78y_male_ICU        : num  0.12 0 0.08 0.01 0 0 0 10.5 2.1 0 ...
##  $ COVID_09_64y_female_ICU      : num  0.51 0.01 0.88 0.02 0.79 ...
##  $ COVID_10_62y_male_ICU        : num  0.1 0 0.13 0.01 0.15 ...
##  $ COVID_11_52y_female_NonICU   : num  0.38 0.02 0.47 0.03 0.08 ...
##  $ COVID_12_50y_male_ICU        : num  0.45 0 0.16 0 1.75 0 0 16 3.61 0 ...
##  $ COVID_13_37y_male_NonICU     : num  0.18 0 0.07 0.01 0 0 0 22.1 2.73 0 ...
##  $ COVID_14_55y_male_ICU        : num  0.23 0 0.22 0.04 0.93 0 0.07 10.3 2.16 0 ...
##  $ COVID_15_68y_male_ICU        : num  0.42 0 0.07 0 0.15 0.03 0 9.37 2.94 0 ...
##  $ COVID_16_48y_male_NonICU     : num  0.41 0.01 0.58 0 0.19 ...
##  $ COVID_17_54y_male_NonICU     : num  0.63 0.02 0.15 0.02 0 ...
##  $ COVID_18_70y_female_NonICU   : num  0.47 0 0.3 0.02 0.06 ...
##  $ COVID_19_51y_male_NonICU     : num  0.33 0.02 0.11 0.02 0 ...
##  $ COVID_20_62y_male_ICU        : num  0.32 0 0.07 0 0.22 ...
##  $ COVID_21_66y_male_ICU        : num  0.18 0 0 0 0.37 0.03 0 7.1 1.11 0 ...
##  $ COVID_22_43y_male_ICU        : num  0.09 0 0.06 0 0.06 0 0.06 5.17 1.05 0 ...
##  $ COVID_23_76y_male_ICU        : num  0.18 0.01 0.03 0 0.07 0.03 0.04 8.87 1.45 0 ...
##  $ COVID_24_55y_male_ICU        : num  0.22 0.01 0.11 0.02 0.15 ...
##  $ COVID_25_55y_male_ICU        : num  0.29 0 0.09 0.03 0 ...
##  $ COVID_26_41y_female_ICU      : num  0.42 0 0.18 0 0.87 ...
##  $ COVID_27_71y_female_ICU      : num  0.16 0.01 0.23 0.01 0.18 ...
##  $ COVID_28_63y_male_ICU        : num  0.18 0 0.18 0.05 0.45 ...
##  $ COVID_29_63y_female_ICU      : num  0.35 0 0.03 0.03 0.15 0.03 0.08 9.74 1.57 0 ...
##  $ COVID_30_54y_male_ICU        : num  0.23 0 0.11 0.01 0 ...
##  $ COVID_31_50y_male_ICU        : num  0.15 0 0.47 0 0 0.03 0 10.4 1.74 0 ...
##  $ COVID_32_72y_male_ICU        : num  0.34 0.01 0.04 0 0.29 0 0.04 8.96 1.88 0 ...
##  $ COVID_33_81y_male_NonICU     : num  0.35 0 0.3 0.06 0.26 ...
##  $ COVID_34_64y_female_NonICU   : num  0.36 0 0.11 0 0.12 ...
##  $ COVID_35_58y_female_NonICU   : num  0.26 0 0.51 0.02 0.16 ...
##  $ COVID_36_68y_male_NonICU     : num  0.18 0.01 0.09 0 0.08 ...
##  $ COVID_37_87y_male_NonICU     : num  0.2 0 0.09 0.07 0.31 ...
##  $ COVID_38_68y_male_ICU        : num  0.29 0 0.1 0.02 0.35 ...
##  $ COVID_39_80y_female_ICU      : num  0.19 0 0.27 0 0 ...
##  $ COVID_40_66y_male_ICU        : num  0.22 0 0.17 0 0.08 0 0 14.6 2.47 0 ...
##  $ COVID_41_74y_male_ICU        : num  0.19 0 0.14 0 0.19 0 0 6.63 1.21 0 ...
##  $ COVID_42_21y_female_ICU      : num  0.24 0.01 0.33 0.01 0.39 0 0 15.1 2.23 0 ...
##  $ COVID_43_83y_female_ICU      : num  0.29 0 0 0 0.11 0 0 5.78 1.44 0 ...
##  $ COVID_44_46y_male_ICU        : num  0.22 0 0.14 0 0 0.04 0 10.8 2.03 0 ...
##  $ COVID_45_62y_female_ICU      : num  0.14 0 0.15 0.03 0.19 0 0 5.36 1.26 0 ...
##  $ COVID_46_62y_male_ICU        : num  0.53 0.01 0.1 0 0.06 ...
##  $ COVID_47_78y_male_ICU        : num  0.08 0.01 0.04 0.03 0.6 ...
##  $ COVID_48_72y_female_ICU      : num  0.19 0 0.06 0.01 0.23 ...
##  $ COVID_49_73y_male_ICU        : num  0.48 0 0.09 0.03 0 ...
##  $ COVID_50_37y_male_ICU        : num  0.08 0 0.01 0 0 0.72 0 6.16 0.62 0 ...
##  $ COVID_51_58y_female_NonICU   : num  0.21 0 0.13 0 0 ...
##  $ COVID_52_71y_male_NonICU     : num  0.25 0.01 0 0.03 0 ...
##  $ COVID_53_35y_female_NonICU   : num  0.25 0 0.64 0.1 0 ...
##  $ COVID_55_62y_female_ICU      : num  0.09 0 0.09 0.01 0 ...
##  $ COVID_56_33y_female_NonICU   : num  0.28 0 0.16 0.09 0.23 ...
##  $ COVID_57_30y_female_NonICU   : num  0.42 0 0.27 0.01 0.19 ...
##  $ COVID_58_62y_male_NonICU     : num  0.39 0 0.08 0 0 ...
##  $ COVID_59_55y_male_NonICU     : num  0.33 0 0.1 0 0.07 ...
##  $ COVID_60_49y_male_NonICU     : num  0.22 0 0.14 0 0 ...
##  $ COVID_61_54y_female_NonICU   : num  0.25 0 0.1 0.03 0.13 0 0 19.8 3.67 0 ...
##  $ COVID_62_78y_female_ICU      : num  0.21 0 0.04 0 0.05 ...
##  $ COVID_63_39y_female_ICU      : num  0.29 0 0.01 0 0.14 ...
##  $ COVID_64_65y_male_ICU        : num  0.38 0.01 0.04 0.02 0.56 0 0.04 9.99 2.14 0 ...
##  $ COVID_65_84y_male_NonICU     : num  0.4 0.01 0.07 0 0.58 ...
##  $ COVID_66_66y_female_NonICU   : num  0.64 0 0 0 0 ...
##  $ COVID_67_57y_male_ICU        : num  0.37 0 0.35 0 0 ...
##  $ COVID_68_79y_male_ICU        : num  0.58 0 0.15 0.01 0 ...
##  $ COVID_69_77y_female_NonICU   : num  0.52 0 0.29 0.02 0 0 0 23.4 4.18 0 ...
##  $ COVID_70_81y_male_NonICU     : num  0.27 0 0.07 0 0 ...
##  $ COVID_71_37y_male_ICU        : num  0.07 0.01 0.12 0.01 0 ...
##  $ COVID_72_50y_female_NonICU   : num  0.52 0 0.1 0.01 0 ...
##  $ COVID_73_82y_male_NonICU     : num  0.46 0.01 0.02 0.02 0.17 ...
##  $ COVID_74_55y_female_ICU      : num  0.24 0 0.12 0.02 0.26 ...
##  $ COVID_75_55y_male_NonICU     : num  0.23 0.01 0.14 0 0 ...
##  $ COVID_76_73y_female_ICU      : num  0.17 0 0.09 0.01 0.04 0 0.04 7.88 0.83 0 ...
##  $ COVID_77_55y_female_ICU      : num  0.05 0 0.01 0 0 ...
##  $ COVID_78_80y_male_NonICU     : num  0.19 0 0.2 0 0 ...
##  $ COVID_79_27y_male_NonICU     : num  0.08 0.01 0.03 0 0 ...
##  $ COVID_80_71y_male_ICU        : num  0.28 0 0.05 0 0.05 ...
##  $ COVID_82_67y_male_NonICU     : num  0.39 0.01 0.1 0 0 0 0 17.1 2.31 0 ...
##  $ COVID_83_85y_female_NonICU   : num  0.47 0 0.18 0.05 0 ...
##  $ COVID_84_75y_female_NonICU   : num  0.35 0 0.03 0 0.17 ...
##  $ COVID_85_62y_male_ICU        : num  0.29 0 0.04 0 0 ...
##  $ COVID_86_52y_female_NonICU   : num  0.6 0 0.27 0.02 0 ...
##  $ COVID_87_61y_male_ICU        : num  0.65 0 0.15 0 0 ...
##  $ COVID_89_90y_female_NonICU   : num  0.2 0 0.07 0.03 0.14 0 0 14.8 1.67 0 ...
##  $ COVID_90_86y_female_NonICU   : num  0.4 0 0.05 0.01 0.31 ...
##  $ COVID_91_29y_female_NonICU   : num  0.6 0 0.03 0.02 0.05 ...
##  $ COVID_92_82y_female_ICU      : num  0.34 0 0.02 0.04 0.58 ...
##  $ COVID_93_81y_female_ICU      : num  0.37 0 0.11 0 0.05 ...
##  $ COVID_94_24y_female_NonICU   : num  0.81 0 0.17 0.02 0 ...
##  $ COVID_95_49y_male_NonICU     : num  0.37 0.01 0.2 0.02 0.15 ...
##  $ COVID_96_51y_male_NonICU     : num  1.61 0 0.02 0 0 ...
##  $ COVID_97_76y_male_ICU        : num  0.19 0 0.02 0.05 0.12 ...
##  $ COVID_98_81y_male_NonICU     : num  0.78 0 0.26 0 0.37 ...
##  $ COVID_99_71y_male_ICU        : num  0.33 0 0.02 0 0.04 0 0 9.76 1.11 0 ...
##  $ COVID_100_74y_female_NonICU  : num  0.3 0 0.09 0 0.04 0 0.02 18.4 1.84 0 ...
##  $ COVID_101_58y_male_ICU       : num  0.33 0 0.11 0.03 0.05 ...
##  $ COVID_102_84y_male_NonICU    : num  0.12 0 0.01 0.01 0 0.07 0 9.1 1.06 0 ...
##   [list output truncated]
```

```
#Load metadata for participants
series <- read.csv(file = "QBS103_GSE157103_series_matrix-1.csv")
dim(series)
```

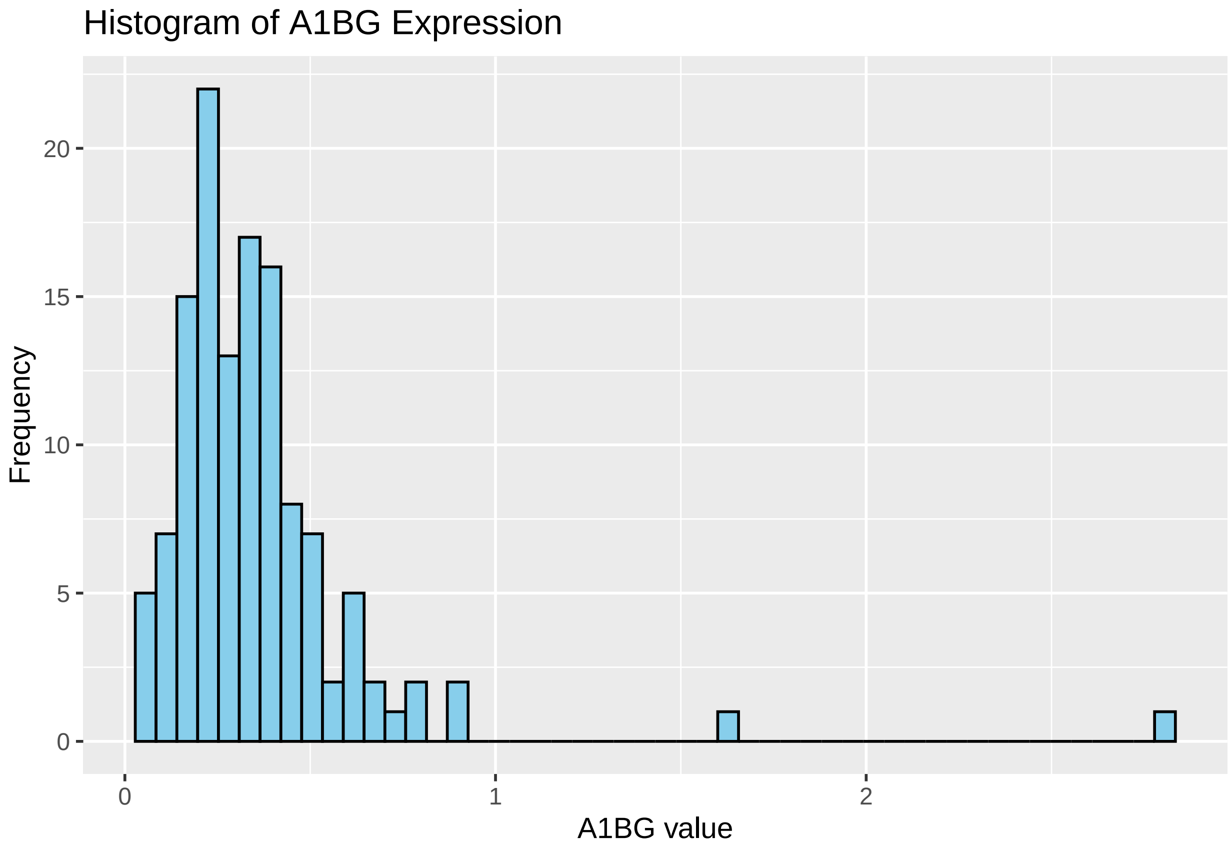
```
## [1] 126  25
```

```
str(series)
```

```
## 'data.frame':    126 obs. of  25 variables:
##  $ participant_id                        : chr  "COVID_01_39y_male_NonICU" "COVID_02_63y_male_NonICU
##  $ geo_accession                         : chr  "GSM4753021" "GSM4753022" "GSM4753023" "GSM4753024"
##  $ status                                : chr  "Public on Aug 29 2020" "Public on Aug 29 2020" "Pu
##  $ X.Sample_submission_date              : chr  "Aug 28 2020" "Aug 28 2020" "Aug 28 2020" "Aug 28 2
##  $ last_update_date                      : chr  "Aug 29 2020" "Aug 29 2020" "Aug 29 2020" "Aug 29 2
##  $ type                                  : chr  "SRA" "SRA" "SRA" "SRA" ...
##  $ channel_count                         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ source_name_ch1                       : chr  "Leukocytes from whole blood" "Leukocytes from whole
##  $ organism_ch1                          : chr  "Homo sapiens" "Homo sapiens" "Homo sapiens" "Homo
##  $ disease_status                        : chr  "disease state: COVID-19" "disease state: COVID-19"
##  $ age                                   : chr  "39" "63" "33" "49" ...
##  $ sex                                   : chr  " male" " male" " male" " male" ...
##  $ icu_status                            : chr  " no" " no" " no" " no" ...
##  $ apacheii                              : chr  "15" " unknown" " unknown" " unknown" ...
##  $ charlson_score                        : int  0 2 2 1 1 1 7 7 2 1 ...
##  $ mechanical_ventilation                : chr  " yes" " no" " no" " no" ...
##  $ ventilator.free_days                  : int  0 28 28 28 23 28 28 0 0 2 ...
##  $ hospital.free_days_post_45_day_followup: int  0 39 18 39 27 36 42 0 0 0 ...
##  $ ferritin.ng.ml.                       : chr  "946" "1060" "1335" "583" ...
##  $ crp.mg.l.                             : chr  "73.1" " unknown" "53.2" "251.1" ...
##  $ ddimer.mg.l_feu.                      : chr  "1.3" "1.03" "1.48" "1.32" ...
##  $ procalcitonin.ng.ml..                  : chr  "36" "0.37" "0.07" "0.98" ...
##  $ lactate.mmol.l.                        : chr  "0.9" " unknown" " unknown" "0.87" ...
##  $ fibrinogen                             : chr  "513" "unknown" "513" "949" ...
##  $ sofa                                   : chr  "8" " unknown" " unknown" " unknown" ...
```

```
#1.
#Histogram of gene expression
#Select the first gene (A1BG) for analysis and convert to numeric vector
new_gene <- as.numeric(gene_data[1, ])
new_gene <- data.frame(value = new_gene)

#Plot histogram for A1BG expression values across all participants
ggplot(new_gene, aes(x = value)) +
  geom_histogram(bins = 50, color = "black", fill = "skyblue") +
  labs(title = "Histogram of A1BG Expression",
       x = "A1BG value",
       y = "Frequency")
```



```
#2.
#Scatterplot of gene expression vs age
#Pivot gene expression row into long format for merging
gene_line1 <- gene_data[1, ] %>%
  pivot_longer(cols = everything(),names_to = "participant_id",values_to = "A1BG_value")
dim(gene_line1)
```

```
## [1] 126   2
```

```r
dim(series)
```

```
## [1] 126  25
```

```r
# Merge with metadata by participant_id to obtain age and other covariates
new_df<-merge(series,gene_line1,by="participant_id")
dim(new_df)
```

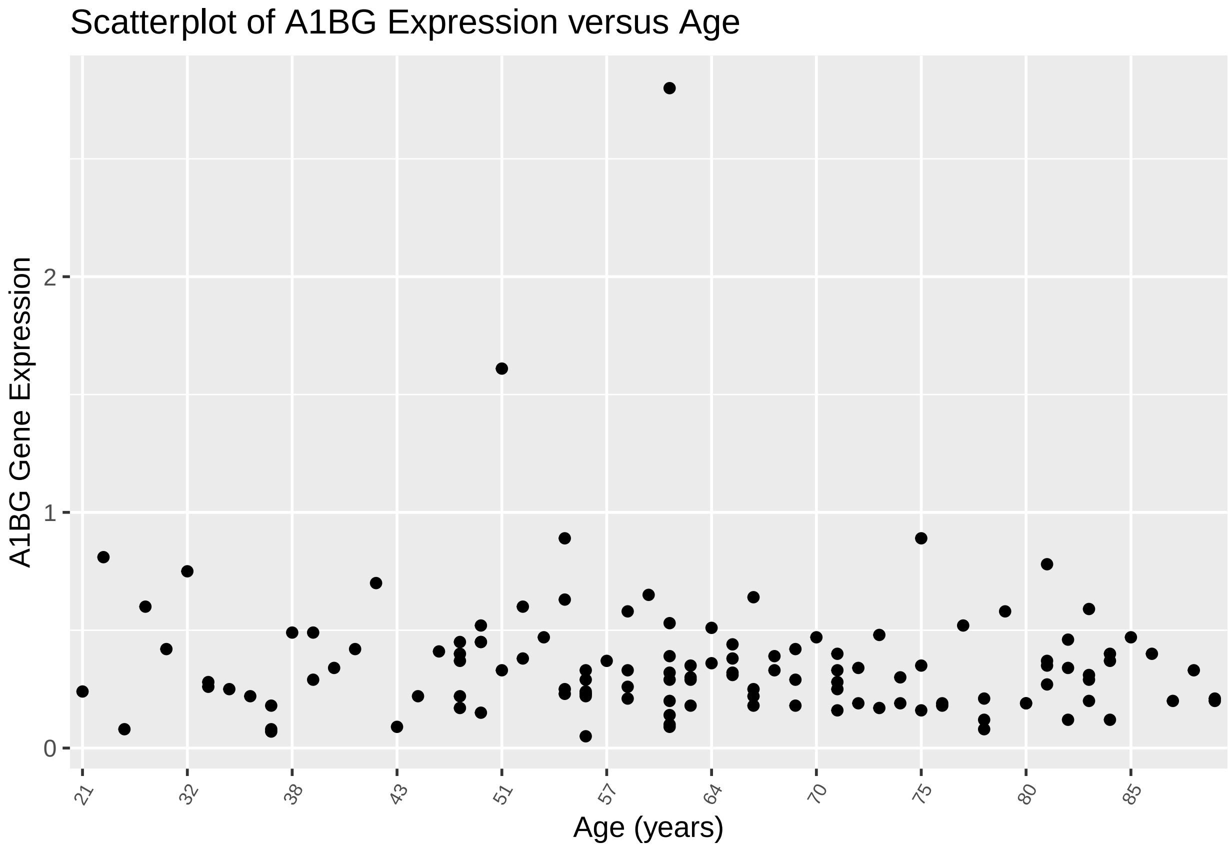
```
## [1] 125  26
```

```r
str(new_df)
```

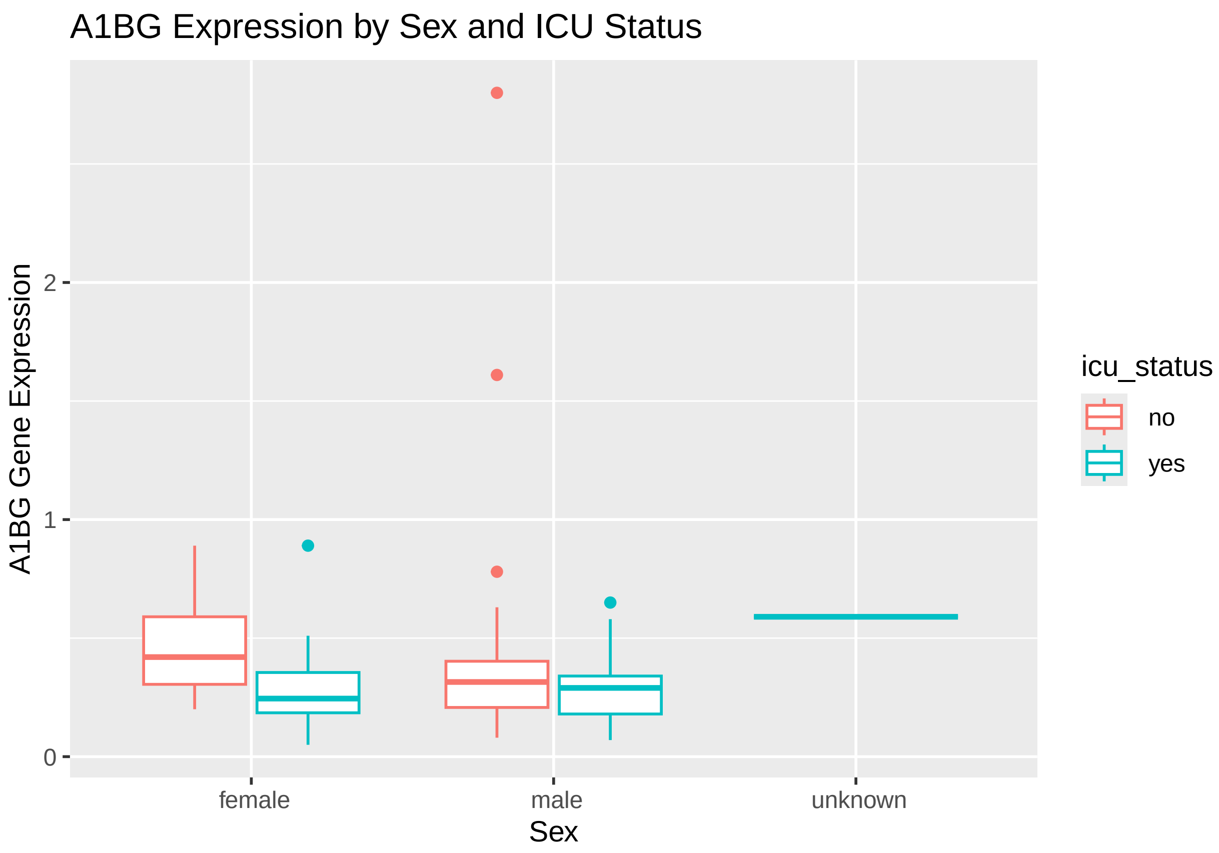
```
## ’data.frame':    125 obs. of  26 variables:
##  $ participant_id                         : chr  "COVID_01_39y_male_NonICU" "COVID_02_63y_male_NonICU
##  $ geo_accession                          : chr  "GSM4753021" "GSM4753022" "GSM4753023" "GSM4753024"
##  $ status                                 : chr  "Public on Aug 29 2020" "Public on Aug 29 2020" "Pub
##  $ X.Sample_submission_date               : chr  "Aug 28 2020" "Aug 28 2020" "Aug 28 2020" "Aug 28 20
##  $ last_update_date                       : chr  "Aug 29 2020" "Aug 29 2020" "Aug 29 2020" "Aug 29 20
##  $ type                                   : chr  "SRA" "SRA" "SRA" "SRA" ...
##  $ channel_count                          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ source_name_ch1                        : chr  "Leukocytes from whole blood" "Leukocytes from whole
##  $ organism_ch1                           : chr  "Homo sapiens" "Homo sapiens" "Homo sapiens" "Homo s
##  $ disease_status                         : chr  "disease state: COVID-19" "disease state: COVID-19"
##  $ age                                    : chr  "39" "63" "33" "49" ...
##  $ sex                                    : chr  " male" " male" " male" " male" ...
##  $ icu_status                             : chr  " no" " no" " no" " no" ...
##  $ apacheii                               : chr  "15" " unknown" " unknown" " unknown" ...
##  $ charlson_score                         : int  0 2 2 1 1 7 7 2 1 6 ...
##  $ mechanical_ventilation                 : chr  " yes" " no" " no" " no" ...
##  $ ventilator.free_days                   : int  0 28 28 28 23 28 0 0 2 28 ...
##  $ hospital.free_days_post_45_day_followup: int  0 39 18 39 27 42 0 0 0 35 ...
##  $ ferritin.ng.ml.                        : chr  "946" "1060" "1335" "583" ...
##  $ crp.mg.l.                              : chr  "73.1" " unknown" "53.2" "251.1" ...
##  $ ddimer.mg.l_feu.                       : chr  "1.3" "1.03" "1.48" "1.32" ...
##  $ procalcitonin.ng.ml..                  : chr  "36" "0.37" "0.07" "0.98" ...
##  $ lactate.mmol.l.                        : chr  "0.9" " unknown" " unknown" "0.87" ...
##  $ fibrinogen                             : chr  "513" "unknown" "513" "949" ...
##  $ sofa                                   : chr  "8" " unknown" " unknown" " unknown" ...
##  $ A1BG_value                             : num  0.49 0.29 0.26 0.45 0.17 0.49 0.12 0.51 0.1 0.3 ...
```

```r
#Prepare age levels for the x-axis: numeric ages in ascending order, then special categories
num_ages <- sort(as.numeric(unique(new_df$age)[!grepl("[^0-9]", unique(new_df$age))]))
special_ages <- unique(new_df$age)[grepl("[^0-9]", unique(new_df$age))]
age_levels <- c(as.character(num_ages), special_ages)

# Scatterplot: A1BG expression vs age (x-axis shows all ages in specified order)
ggplot(new_df,aes(x=factor(age, levels = age_levels),y=A1BG_value))+
  geom_point()+
  labs(title = "Scatterplot of A1BG Expression versus Age",
       x = "Age (years)",
       y = "A1BG Gene Expression")+
  theme(axis.text.x = element_text(angle = 60, hjust = 1, size = 7))+
  scale_x_discrete(breaks = age_levels[seq(1, length(age_levels), by = 5)])
```



```
#3.
#Boxplot of gene expression by sex and ICU status
#sex: categorical variable; icu_status: categorical variable
ggplot(new_df,aes(x=sex,y=A1BG_value,color=icu_status))+
  geom_boxplot()+
  labs(
    title = "A1BG Expression by Sex and ICU Status",
    x = "Sex",
    y = "A1BG Gene Expression",
    fill = "ICU Status")
```


A1BG Expression by Sex and ICU Status
A1BG Gene Expression
0
1
2
female
male
unknown
Sex
icu_status
no
yes