submission1_XinyiYu

2025-07-15

Identify one gene, one continuous covariate, and two categorical covariates in the provided dataset. Gene : AAAS continuous covariate: age categorical covariates: sex & disease_status

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.1
## v purrr     1.0.1
## -- 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
```

```
#read data
gene_data <- read.csv("/Users/yuxinyi/Dartmouth/Data Science/QBS103_GSE157103_genes.csv", row.names = 1)
meta_data <- read.csv("/Users/yuxinyi/Dartmouth/Data Science/QBS103_GSE157103_series_matrix-1.csv", row
#select gene
aaas <- gene_data["AAAS", ]  # returns a named vector
#convert to data frame
aaas_df <- data.frame(SampleID = names(aaas),
                      AAAS_expression = as.numeric(aaas),
                      row.names = NULL)
head(aaas_df)
```

```
##                   SampleID AAAS_expression
## 1 COVID_01_39y_male_NonICU           18.92
## 2 COVID_02_63y_male_NonICU           18.68
## 3 COVID_03_33y_male_NonICU           13.85
## 4 COVID_04_49y_male_NonICU           22.11
## 5 COVID_05_49y_male_NonICU            8.45
## 6  COVID_06_.y_male_NonICU           19.60
```

```
head(meta_data)
```

```
##                          geo_accession                status
## COVID_01_39y_male_NonICU    GSM4753021 Public on Aug 29 2020
## COVID_02_63y_male_NonICU    GSM4753022 Public on Aug 29 2020
## COVID_03_33y_male_NonICU    GSM4753023 Public on Aug 29 2020
## COVID_04_49y_male_NonICU    GSM4753024 Public on Aug 29 2020
```

```
## COVID_05_49y_male_NonICU    GSM4753025 Public on Aug 29 2020
## COVID_06_:y_male_NonICU     GSM4753026 Public on Aug 29 2020
##                          X.Sample_submission_date last_update_date type
## COVID_01_39y_male_NonICU              Aug 28 2020      Aug 29 2020  SRA
## COVID_02_63y_male_NonICU              Aug 28 2020      Aug 29 2020  SRA
## COVID_03_33y_male_NonICU              Aug 28 2020      Aug 29 2020  SRA
## COVID_04_49y_male_NonICU              Aug 28 2020      Aug 29 2020  SRA
## COVID_05_49y_male_NonICU              Aug 28 2020      Aug 29 2020  SRA
## COVID_06_:y_male_NonICU               Aug 28 2020      Aug 29 2020  SRA
##                          channel_count             source_name_ch1 organism_ch1
## COVID_01_39y_male_NonICU             1 Leukocytes from whole blood Homo sapiens
## COVID_02_63y_male_NonICU             1 Leukocytes from whole blood Homo sapiens
## COVID_03_33y_male_NonICU             1 Leukocytes from whole blood Homo sapiens
## COVID_04_49y_male_NonICU             1 Leukocytes from whole blood Homo sapiens
## COVID_05_49y_male_NonICU             1 Leukocytes from whole blood Homo sapiens
## COVID_06_:y_male_NonICU              1 Leukocytes from whole blood Homo sapiens
##                                    disease_status age   sex icu_status apacheii
## COVID_01_39y_male_NonICU disease state: COVID-19  39  male         no       15
## COVID_02_63y_male_NonICU disease state: COVID-19  63  male         no  unknown
## COVID_03_33y_male_NonICU disease state: COVID-19  33  male         no  unknown
## COVID_04_49y_male_NonICU disease state: COVID-19  49  male         no  unknown
## COVID_05_49y_male_NonICU disease state: COVID-19  49  male         no       19
## COVID_06_:y_male_NonICU  disease state: COVID-19   :  male         no  unknown
##                          charlson_score mechanical_ventilation
## COVID_01_39y_male_NonICU              0                    yes
## COVID_02_63y_male_NonICU              2                     no
## COVID_03_33y_male_NonICU              2                     no
## COVID_04_49y_male_NonICU              1                     no
## COVID_05_49y_male_NonICU              1                    yes
## COVID_06_:y_male_NonICU               1                     no
##                          ventilator.free_days
## COVID_01_39y_male_NonICU                    0
## COVID_02_63y_male_NonICU                   28
## COVID_03_33y_male_NonICU                   28
## COVID_04_49y_male_NonICU                   28
## COVID_05_49y_male_NonICU                   23
## COVID_06_:y_male_NonICU                    28
##                          hospital.free_days_post_45_day_followup
## COVID_01_39y_male_NonICU                                      0
## COVID_02_63y_male_NonICU                                     39
## COVID_03_33y_male_NonICU                                     18
## COVID_04_49y_male_NonICU                                     39
## COVID_05_49y_male_NonICU                                     27
## COVID_06_:y_male_NonICU                                      36
##                          ferritin.ng.ml. crp.mg.l. ddimer.mg.l_feu.
## COVID_01_39y_male_NonICU             946      73.1              1.3
## COVID_02_63y_male_NonICU            1060   unknown             1.03
## COVID_03_33y_male_NonICU            1335      53.2             1.48
## COVID_04_49y_male_NonICU             583     251.1             1.32
## COVID_05_49y_male_NonICU             800     355.8             0.69
## COVID_06_:y_male_NonICU              563     129.1          unknown
##                          procalcitonin.ng.ml.. lactate.mmol.l. fibrinogen
## COVID_01_39y_male_NonICU                    36             0.9        513
## COVID_02_63y_male_NonICU                 0.37         unknown    unknown
```

```
## COVID_03_33y_male_NonICU                 0.07         unknown        513
## COVID_04_49y_male_NonICU                 0.98            0.87        949
## COVID_05_49y_male_NonICU                 4.92            1.48        929
## COVID_06_:y_male_NonICU                  0.67            0.86        769
##                                sofa
## COVID_01_39y_male_NonICU          8
## COVID_02_63y_male_NonICU    unknown
## COVID_03_33y_male_NonICU    unknown
## COVID_04_49y_male_NonICU    unknown
## COVID_05_49y_male_NonICU          7
## COVID_06_:y_male_NonICU     unknown
```

Select variables from metadata

```
variable_df <- meta_data %>%
  rownames_to_column(var = "SampleID") %>% #convert row names to a column named 'SampleID'
  select(SampleID, age, sex, disease_status) %>%
  mutate(
    age = as.numeric(as.character(age)), #convert age to numeric
    sex = as.factor(sex), # convert sex to factor
    disease_status = as.factor(disease_status) #convert disease_status to factor
  )
```

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

```
head(variable_df)
```

```
##                    SampleID age   sex          disease_status
## 1 COVID_01_39y_male_NonICU  39  male disease state: COVID-19
## 2 COVID_02_63y_male_NonICU  63  male disease state: COVID-19
## 3 COVID_03_33y_male_NonICU  33  male disease state: COVID-19
## 4 COVID_04_49y_male_NonICU  49  male disease state: COVID-19
## 5 COVID_05_49y_male_NonICU  49  male disease state: COVID-19
## 6  COVID_06_:y_male_NonICU  NA  male disease state: COVID-19
```

merge two dataframe

```
#merge two dataframe
merged_data_for_plotting <- aaas_df %>%
  left_join(variable_df, by = "SampleID")

#show the merged data
head(merged_data_for_plotting)
```

```
##                    SampleID AAAS_expression age   sex          disease_status
## 1 COVID_01_39y_male_NonICU           18.92  39  male disease state: COVID-19
## 2 COVID_02_63y_male_NonICU           18.68  63  male disease state: COVID-19
## 3 COVID_03_33y_male_NonICU           13.85  33  male disease state: COVID-19
## 4 COVID_04_49y_male_NonICU           22.11  49  male disease state: COVID-19
## 5 COVID_05_49y_male_NonICU            8.45  49  male disease state: COVID-19
## 6  COVID_06_.y_male_NonICU           19.60  NA  <NA>                   <NA>
```

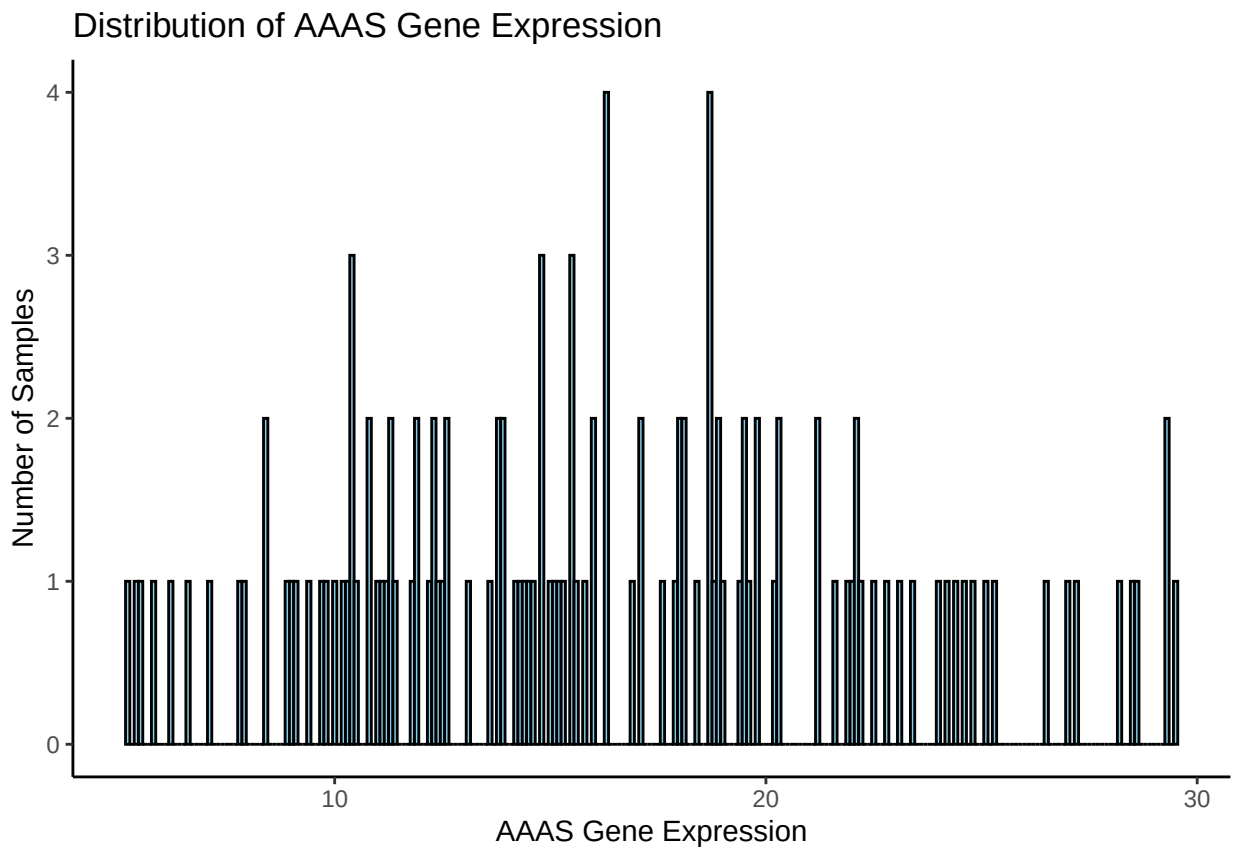
```
summary(merged_data_for_plotting)
```

```
##    SampleID         AAAS_expression      age             sex
##  Length:126         Min.   : 5.17   Min.   :21.00    female :51
##  Class :character   1st Qu.:11.48   1st Qu.:50.50    male   :73
##  Mode  :character   Median :15.57   Median :62.00    unknown: 1
##                     Mean   :16.24   Mean   :61.24   NA's    : 1
##                     3rd Qu.:19.79   3rd Qu.:74.00
##                     Max.   :29.46   Max.   :88.00
##                                     NA's   :3
##                      disease_status
##  disease state: COVID-19    :99
##  disease state: non-COVID-19:26
##  NA's                       : 1
##
##
##
##
```

Histogram for gene expression

```
library(ggplot2)
## set clean and minimal theme
newBlankTheme <- theme(
        panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black", linewidth = rel(1)),
        plot.background = element_rect(fill = "white"),
        panel.background = element_blank(),
        legend.key = element_rect(fill = 'white'),
        legend.position = 'top')

# Histogram of AAAS expression
ggplot(merged_data_for_plotting, aes(x = AAAS_expression)) +
  geom_histogram(binwidth = 0.1, fill = "skyblue", color = "black") +
  labs(title = "Distribution of AAAS Gene Expression",x = "AAAS Gene Expression",y = "Number of Samples
  newBlankTheme
```



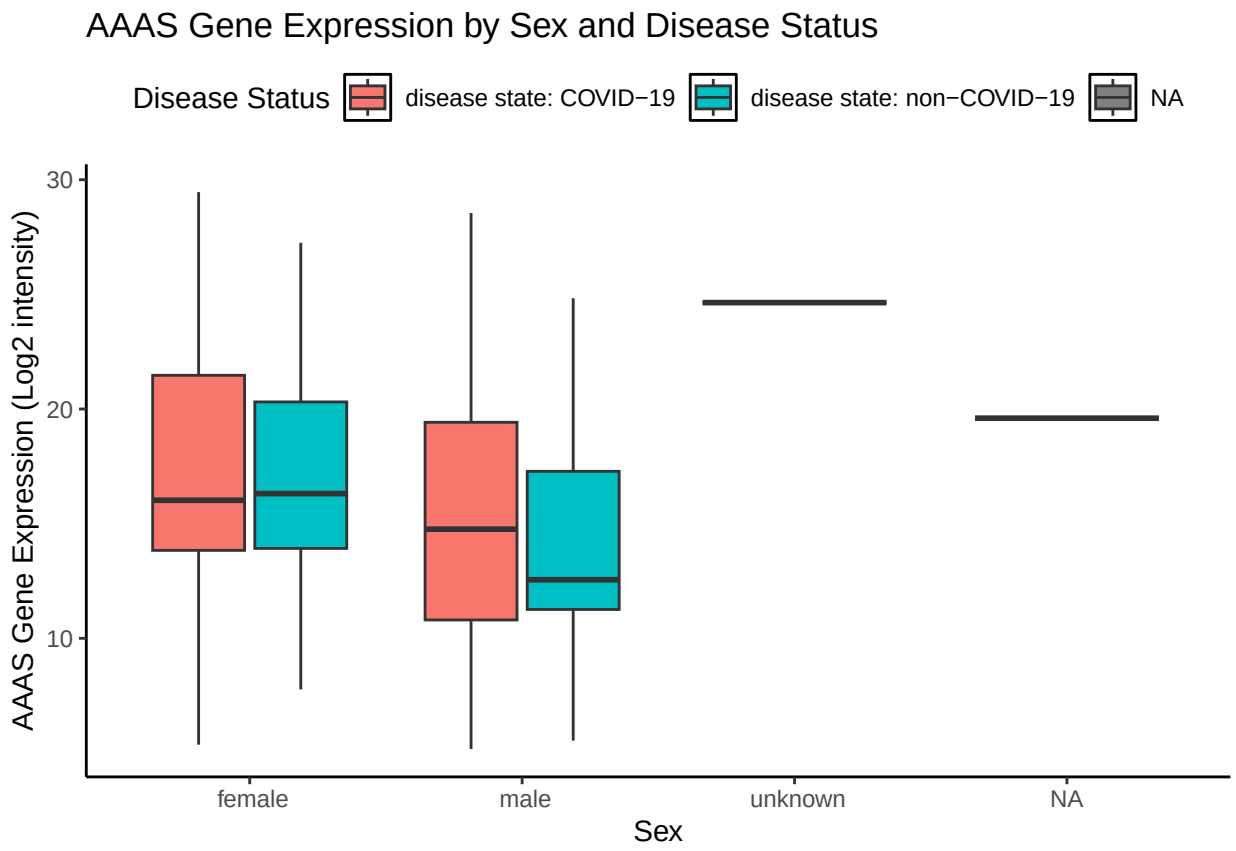
Scatterplot for gene expression and continuous covariate

```
#Scatterplot with Continuous Covariate and gene expression
ggplot(merged_data_for_plotting, aes(x = age, y = AAAS_expression)) +
  geom_point() +
  labs(
    title = "AAAS Gene Expression vs. Age",
    x = "Age (Years)",
    y = "AAAS Gene Expression (Log2 intensity)"
  ) +
  newBlankTheme
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

Boxplot of gene expression separated by both categorical covariates

```
#Boxplot by Categorical Covariates
ggplot(merged_data_for_plotting, aes(x = sex, y = AAAS_expression, fill = disease_status)) +
  geom_boxplot() +
  labs(
    title = "AAAS Gene Expression by Sex and Disease Status",
    x = "Sex",
    y = "AAAS Gene Expression (Log2 intensity)",
    fill = "Disease Status"
  ) +
  newBlankTheme
```


AAAS Gene Expression by Sex and Disease Status
Disease Status
disease state: COVID−19
disease state: non−COVID−19
NA
AAAS Gene Expression (Log2 intensity)
30
20
10
female
male
unknown
NA
Sex