main

2024-07-20

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
## Warning: package 'ggplot2' was built under R version 4.3.3
## Warning: package 'tibble' was built under R version 4.3.3
## Warning: package 'tidyr' was built under R version 4.3.3
## Warning: package 'readr' was built under R version 4.3.3
## Warning: package 'purrr' was built under R version 4.3.3
## Warning: package 'dplyr' was built under R version 4.3.3
## Warning: package 'stringr' was built under R version 4.3.3
## Warning: package 'forcats' was built under R version 4.3.3
## Warning: package 'lubridate' was built under R version 4.3.3
## -- 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
```

```
# import the csv

series_matrix<- read.csv('./data/QBS103_GSE157103_series_matrix.csv')
genes <- read.csv('./data/QBS103_GSE157103_genes.csv')
```

```
head(series_matrix)
```

```
##             participant_id geo_accession                status
## 1 COVID_01_39y_male_NonICU    GSM4753021 Public on Aug 29 2020
## 2 COVID_02_63y_male_NonICU    GSM4753022 Public on Aug 29 2020
## 3 COVID_03_33y_male_NonICU    GSM4753023 Public on Aug 29 2020
## 4 COVID_04_49y_male_NonICU    GSM4753024 Public on Aug 29 2020
## 5 COVID_05_49y_male_NonICU    GSM4753025 Public on Aug 29 2020
## 6  COVID_06_:y_male_NonICU    GSM4753026 Public on Aug 29 2020
##   X.Sample_submission_date last_update_date type channel_count
## 1              Aug 28 2020      Aug 29 2020  SRA             1
## 2              Aug 28 2020      Aug 29 2020  SRA             1
## 3              Aug 28 2020      Aug 29 2020  SRA             1
```

```
## 4              Aug 28 2020      Aug 29 2020  SRA             1
## 5              Aug 28 2020      Aug 29 2020  SRA             1
## 6              Aug 28 2020      Aug 29 2020  SRA             1
##               source_name_ch1 organism_ch1          disease_status age   sex
## 1 Leukocytes from whole blood Homo sapiens disease state: COVID-19  39  male
## 2 Leukocytes from whole blood Homo sapiens disease state: COVID-19  63  male
## 3 Leukocytes from whole blood Homo sapiens disease state: COVID-19  33  male
## 4 Leukocytes from whole blood Homo sapiens disease state: COVID-19  49  male
## 5 Leukocytes from whole blood Homo sapiens disease state: COVID-19  49  male
## 6 Leukocytes from whole blood Homo sapiens disease state: COVID-19   :  male
##   icu_status apacheii charlson_score mechanical_ventilation
## 1         no       15              0                   yes
## 2         no  unknown              2                    no
## 3         no  unknown              2                    no
## 4         no  unknown              1                    no
## 5         no       19              1                   yes
## 6         no  unknown              1                    no
##   ventilator.free_days hospital.free_days_post_45_day_followup ferritin.ng.ml.
## 1                    0                                       0            946
## 2                   28                                      39           1060
## 3                   28                                      18           1335
## 4                   28                                      39            583
## 5                   23                                      27            800
## 6                   28                                      36            563
##   crp.mg.l. ddimer.mg.l_feu. procalcitonin.ng.ml.. lactate.mmol.l. fibrinogen
## 1      73.1             1.3                    36            0.9        513
## 2   unknown            1.03                  0.37        unknown    unknown
## 3      53.2            1.48                  0.07        unknown        513
## 4     251.1            1.32                  0.98           0.87        949
## 5     355.8            0.69                  4.92           1.48        929
## 6     129.1         unknown                  0.67           0.86        769
##       sofa
## 1        8
## 2  unknown
## 3  unknown
## 4  unknown
## 5        7
## 6  unknown
```

```
unique(series_matrix$disease_status) # disease status is one categorical
```

```
## [1] "disease state: COVID-19"     "disease state: non-COVID-19"
```

```
unique(series_matrix$mechanical_ventilation) # another categorial
```

```
## [1] " yes" " no"
```

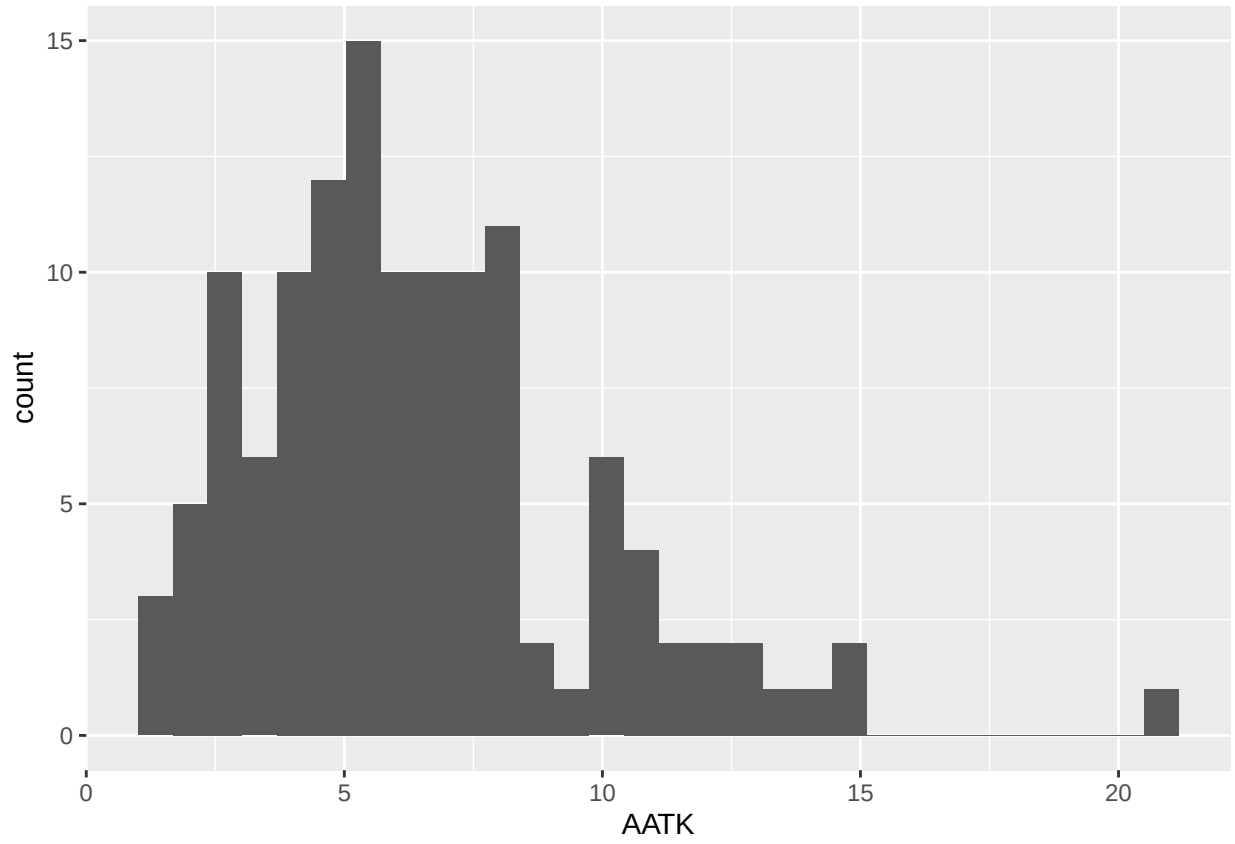
```
unique(series_matrix$age) # this one is covariate
```

```
##  [1] "39"   "63"   "33"   "49"   " :"   "38"   "78"   "64"   "62"   "52"
## [11] "50"   "37"   "55"   "68"   "48"   "54"   "70"   "51"   "66"   "43"
## [21] "76"   "41"   "71"   "72"   "81"   "58"   "87"   "80"   "74"   "21"
## [31] "83"   "46"   "73"   "35"   "30"   "65"   "84"   "57"   "79"   "77"
## [41] "82"   "27"   "67"   "85"   "75"   "61"   " >89" "86"   "29"   "24"
## [51] "53"   "40"   "88"   "42"   "32"   "36"
```

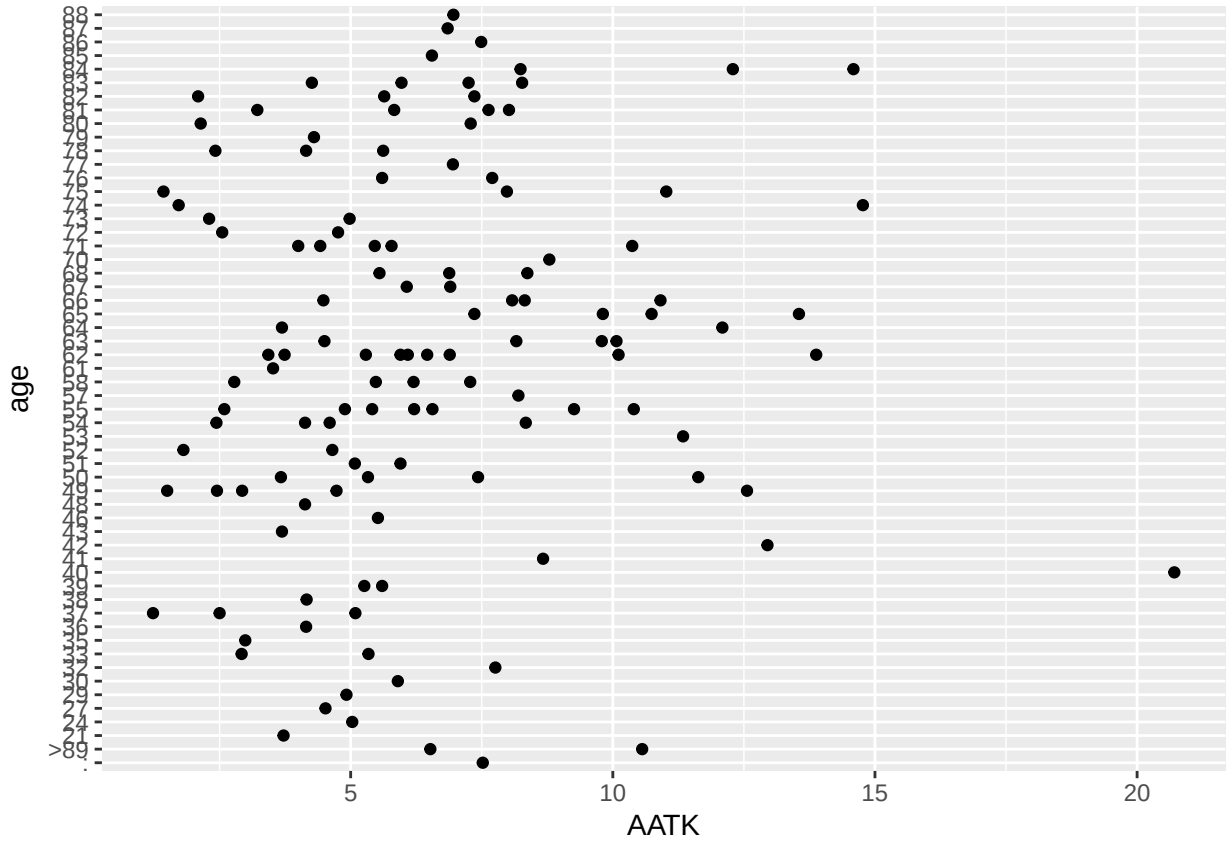
```r
# let's merge these two data sets together
new_df <- series_matrix[c('participant_id', 'disease_status', 'sex', 'age') ]

# get the genes. First column is just the name so slice from 2 onwards
# getting the 29th gene in the row
one_genes <- genes[29, 2:length(genes)]

# transpose and convert to dataframe
temp_df <- as.data.frame(t(one_genes))
colnames(temp_df) <- c('AATK')

# assign the new colunm to the new_df which has the variables we want
new_df$AATK<- temp_df$AATK
```

```r
library(ggplot2)

# creating a histogram
# http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visual

ggplot(new_df, aes(x = AATK)) + geom_histogram()
```

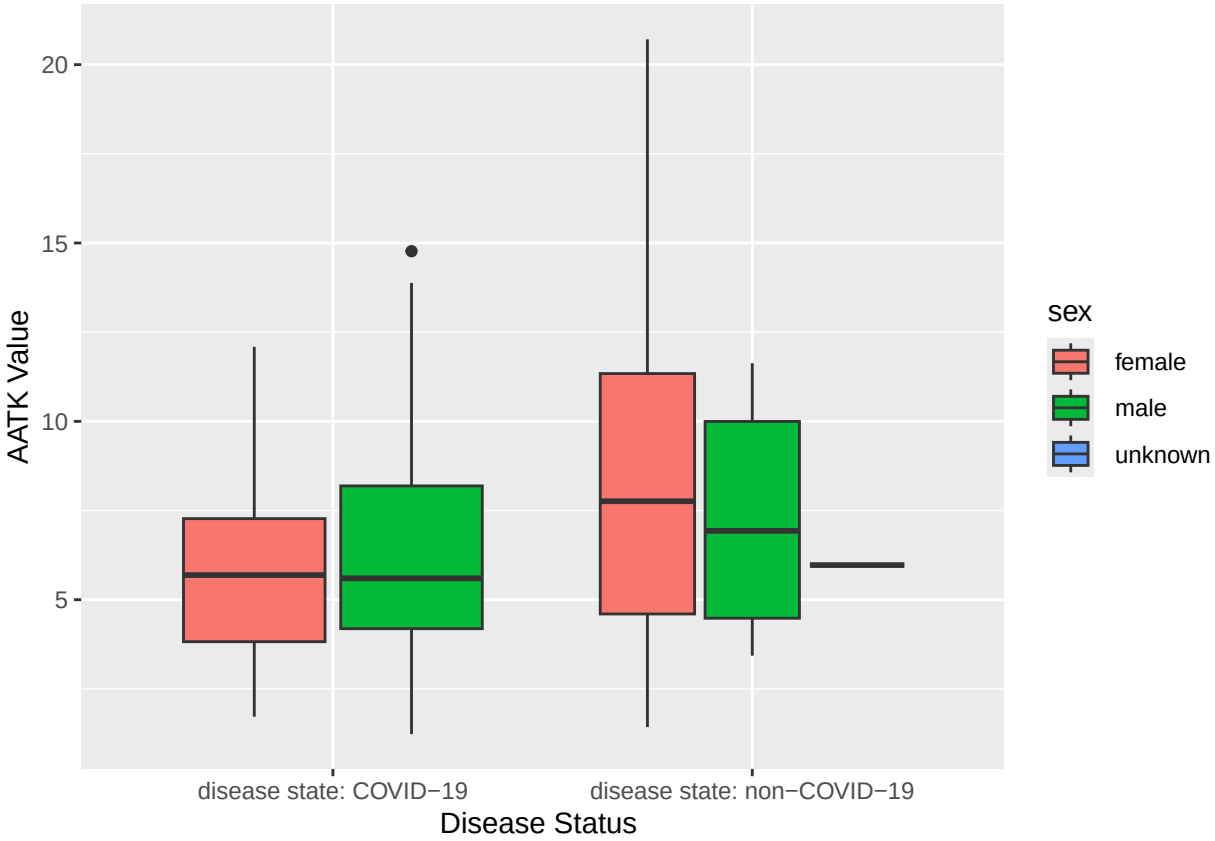
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
# creating a scatter point
ggplot(new_df, aes(x = AATK, y = age)) + geom_point()
```




```
# https://stackoverflow.com/questions/55180015/use-geom-boxplot-with-variable-of-type-double-on-x-axis
# Used this link to plot both box plots

ggplot(new_df, aes(y = AATK, x = disease_status, fill = sex)) +
  geom_boxplot() +
  labs(x = "Disease Status ", y = "AATK Value")
```


20
15
10
5
AATK Value
disease state: COVID−19
disease state: non−COVID−19
Disease Status
sex
female
male
unknown