Project–Submission 1

2024-07-30

Submission 1 (7/30)

1.Identify one gene, one continuous covariate, and two categorical covariates in the provided dataset. Note: Gene expression data and metadata are in two separate files and will need to be linked.

```
#Load packages
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
```

```
#Set the working directory to the path of given dataset.
getwd()
```

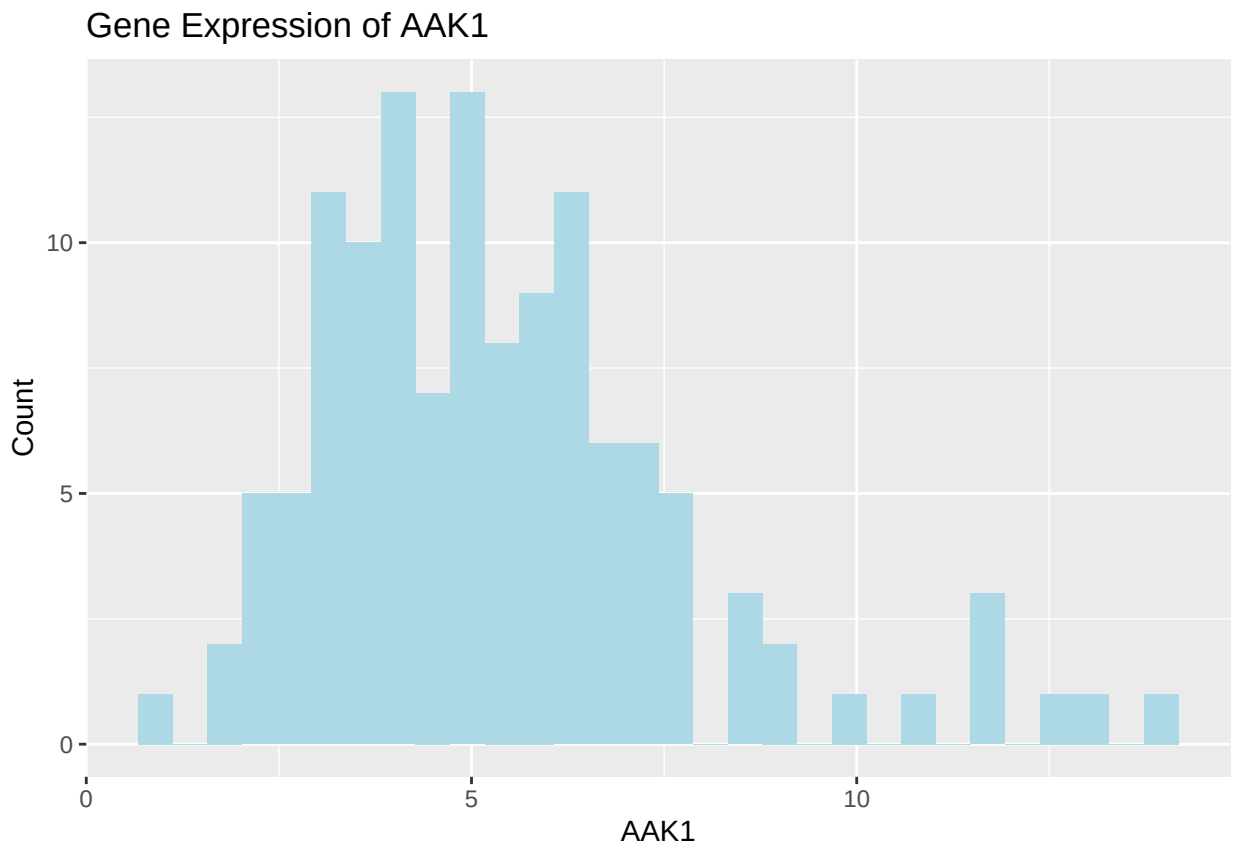
```
## [1] "E:/55555 Dartmouth QBS/QBS 103 Foundation of DS(R)/Project"
```

```
#Using read.csv to read in the gene expression data and metadata
genes_expression<-read.csv(file="QBS103_GSE157103_genes.csv",header=T,stringsAsFactors=F)
metadata<-read.csv(file="QBS103_GSE157103_series_matrix.csv",header=T,stringsAsFactors=F)

#Using 'which' to select the chosen gene and identify it
gene_AAK1<-genes_expression[which(genes_expression[,1]=='AAK1'),]

#use pipe and merge to convert the gene expression to required format and link two dataframes
gene_AAK1<-gene_AAK1 %>%
  gather(key=participant_id,value=expression)
linked_data<-merge(metadata,gene_AAK1)
#head(linked_data)
```

2.Generate the following three plots using ggplot2 for your covariates of choice: 2.1 Histogram for gene expression (5 pts)

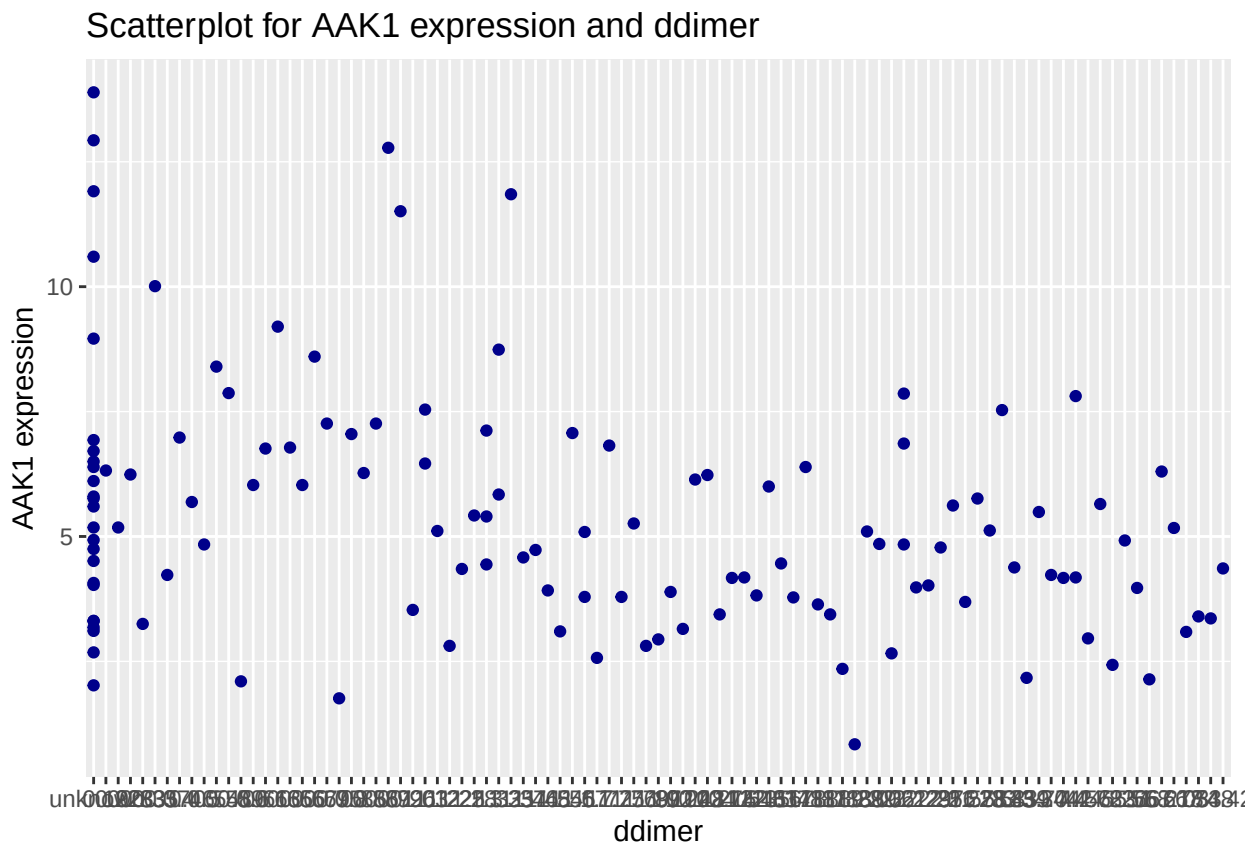
```
ggplot(linked_data,aes(x=as.numeric(expression)))+
  geom_histogram(fill='lightblue')+
  labs(title="Gene Expression of AAK1",
       x="AAK1",
       y="Count")
```

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



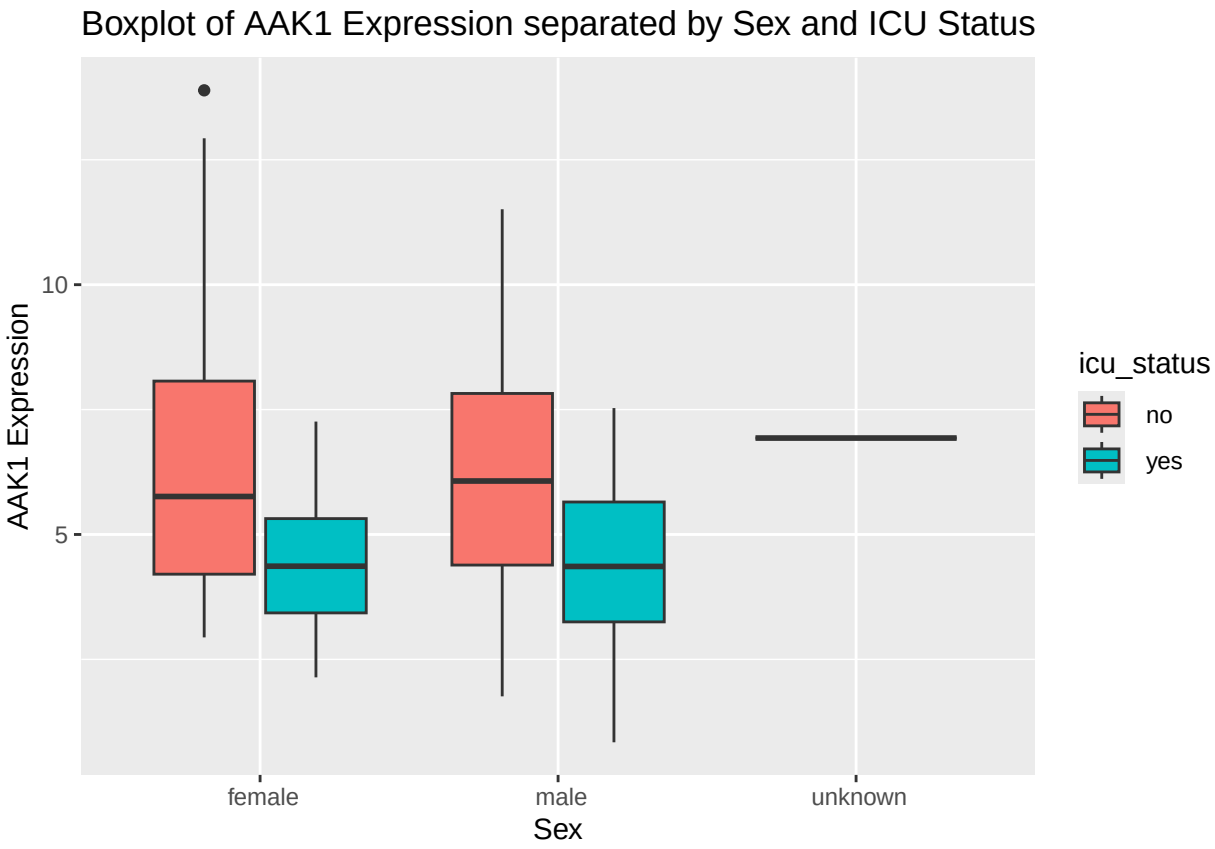
2.2 Scatterplot for gene expression and continuous covariate (5 pts)

```
linked_data$expression<-as.numeric(linked_data$expression)
ggplot(linked_data,aes(x=ddimer.mg.l_feu.,y=expression))+
  geom_point(color='darkblue')+
  labs(title='Scatterplot for AAK1 expression and ddimer',x='ddimer',y='AAK1 expression')
```



2.3 Boxplot of gene expression separated by both categorical covariates (5 pts)

```
ggplot(linked_data,aes(x=sex,y=expression,fill = icu_status))+
  geom_boxplot()+
  labs(title="Boxplot of AAK1 Expression separated by Sex and ICU Status",
       x="Sex",
       y="AAK1 Expression")
```


Boxplot of AAK1 Expression separated by Sex and ICU Status
AAK1 Expression
10
5
female
male
unknown
Sex
icu_status
no
yes