Submission_1_103

R Markdown

Create a git repository for your project and push at least once prior to the first presentation with all the code you are presenting in class. See grading breakdown for final submission and bonus for details. 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. Generate the following three plots using ggplot2 for your covariates of choice: Histogram for gene expression (5 pts)
Scatterplot for gene expression and continuous covariate (5 pts) Boxplot of gene expression separated by both categorical covariates (5 pts) Present your scatterplot in class. Be prepared to explain the gene and covariate you chose and comment on the distribution as if you were presenting your research findings. No slides are required, just bring your plot. In class, be prepared to provide constructive feedback for your classmates (5 pts) Submit your clearly commented code and generated plots as a knitted R markdown file.

```
#Convert genes into gene series

library(readr)
setwd("/Users/deepthi/Documents/GitHub/Final_103_Project_LDG")
# read in genes file
Genes<- read_csv("QBS103_GSE157103_genes.csv")
```

```
## New names:
## Rows: 100 Columns: 127
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): ...1 dbl (126): COVID_01_39y_male_NonICU, COVID_02_63y_male_NonICU,
## COVID_03_33y_...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
#read in Gene Series file
Genes_Series <- read_csv("QBS103_GSE157103_series_matrix.csv")
```

```
## Rows: 126 Columns: 25
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (21): participant_id, geo_accession, status, !Sample_submission_date, la...
## dbl  (4): channel_count, charlson_score, ventilator-free_days, hospital-free...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#transpose gene data to put participant id as rows and genes as columns
TGenes <-as.data.frame(t(Genes))
# rearrange table to make gene names as column names instead of row names
names(TGenes) <- TGenes[1,]
#remove repetitive first column with genes
TGenes <- TGenes[-1,]
#create a column with the row names( participant ids)to merge with gene series file
TGenes$participant_id <- row.names(TGenes)
# Merge Gene and Gene Series tables by participant_id
MergedGenes <- merge(TGenes,Genes_Series, by = 'participant_id', all = TRUE)
```

```r
#load in dplyr to filter table
library(dplyr)
```

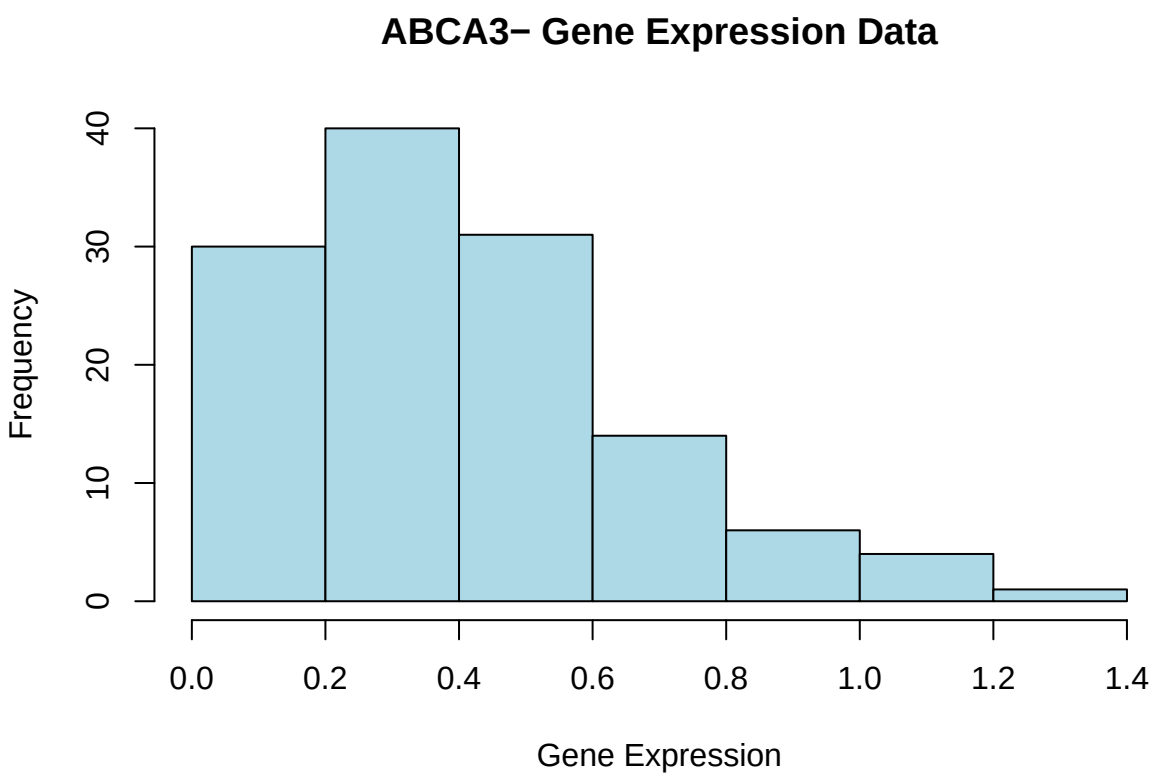
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
#pull out ABCA3 gene expression data, Age data, and ICU_Data information and save into new data frame
Genes_CoVariate_Data <- MergedGenes %>%
  select(ABCA3,age,sex,icu_status)
```

```r
#convert gene expression to numeric data points
Genes_CoVariate_Data$ABCA3 <- as.numeric(Genes_CoVariate_Data$ABCA3)
# plot histogram of gene expression data
hist(Genes_CoVariate_Data$ABCA3,main = 'ABCA3- Gene Expression Data',xlab = 'Gene Expression', col = "li
```

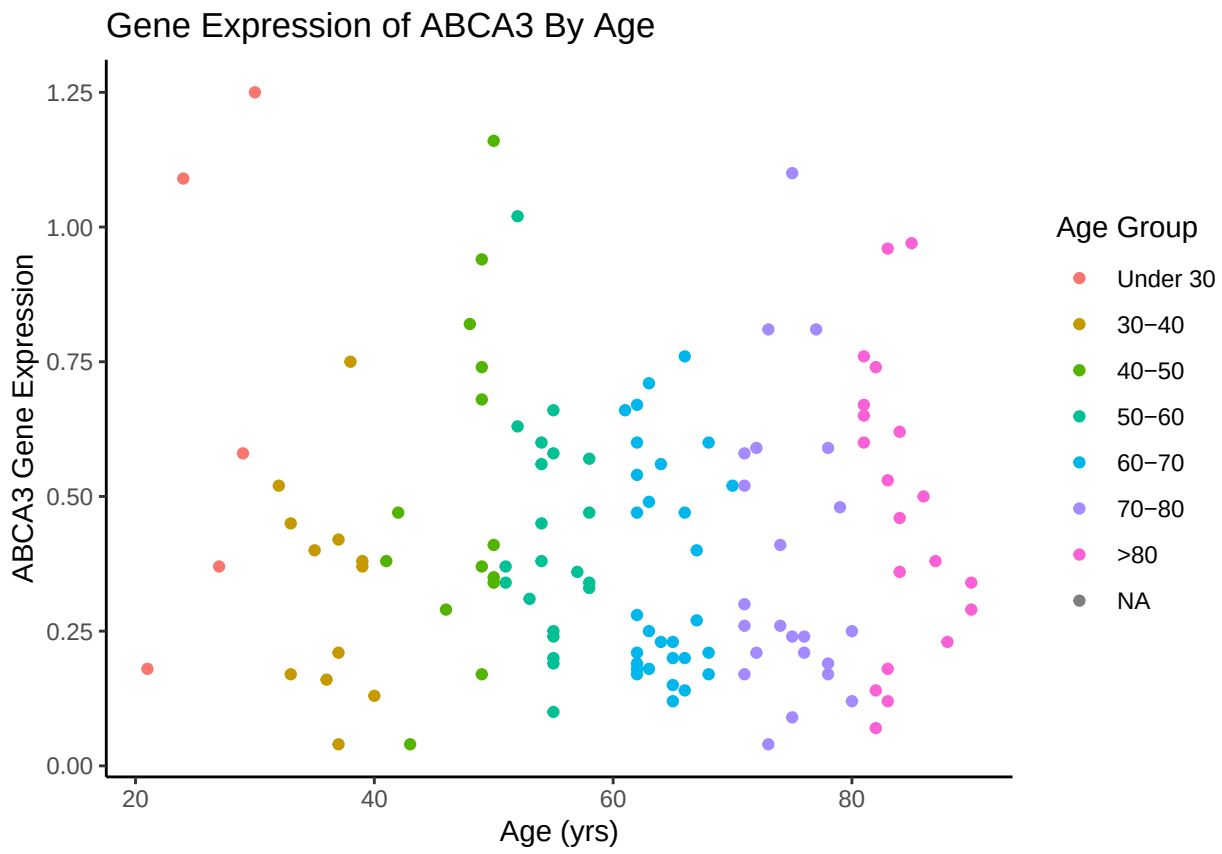



```
#Scatterplot for gene expression and continuous covariate (5 pts)
# ABCA3 is gene provides instructions for making a protein involved in lung surfactant production. Surf

# replace over >89 data to 90 in order to have all age values as numeric
Genes_CoVariate_Data$age[Genes_CoVariate_Data$age== ">89"] <- 90
#convert age col to numeric data
Genes_CoVariate_Data$age <- as.numeric(Genes_CoVariate_Data$age)
```

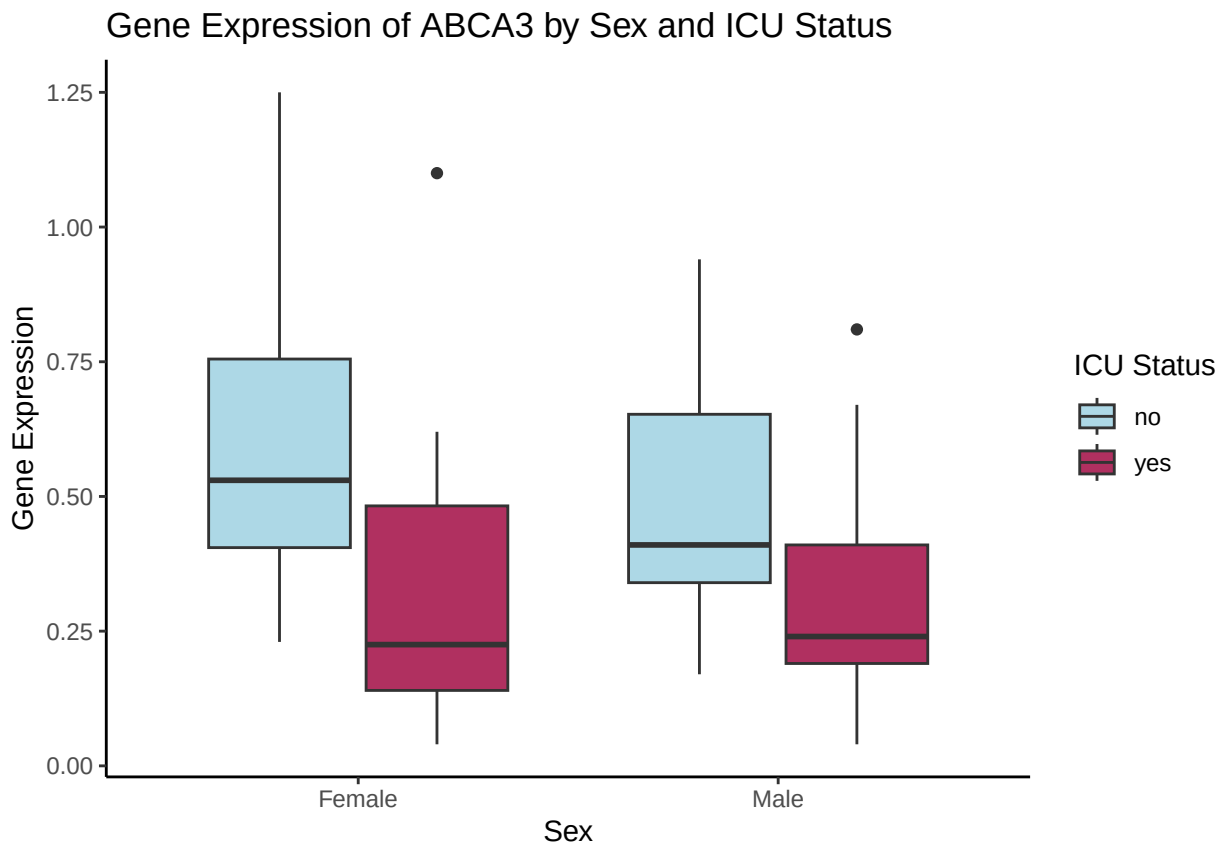
```
## Warning: NAs introduced by coercion
```

```
# load ggplot package
library(ggplot2)
# create age groups based on agevalues
Genes_CoVariate_Data$AgeGroup <- cut(Genes_CoVariate_Data$age,breaks = c(0,30,40,50,60,70,80,90),
                          labels = c('Under 30','30-40','40-50','50-60','60-70','70-80','>80'))

# plot scatter plot of Gene Expression of ABCA3 By Age
ggplot(Genes_CoVariate_Data,aes(x= age,y= ABCA3, color= AgeGroup)) +
  geom_point() +
  labs(title= 'Gene Expression of ABCA3 By Age' ,x = 'Age (yrs)',y = 'ABCA3 Gene Expression', color = "
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```




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

# remove unknown values from data set
Genes_CoVariate_Data[Genes_CoVariate_Data== "unknown"] <- NA
Genes_CoVariate_Data<- na.omit(Genes_CoVariate_Data)

# add color pallete to color boxplots
colorPalette = c('light blue',' maroon')

# creatre box plots to show Gene Expression of ABCA3 by Sex and ICU Statu
ggplot(Genes_CoVariate_Data, aes(x = sex,y= ABCA3, fill=icu_status)) +
  # Add box plot
  geom_boxplot() +
  # add colors
  scale_fill_manual(values= colorPalette) +
  # change x value labels
  scale_x_discrete(labels = c("Female", "Male"))+
  # Change axis labels
  labs(title ='Gene Expression of ABCA3 by Sex and ICU Status', x = 'Sex',y = 'Gene Expression',fill =
  theme_classic()
```


Gene Expression of ABCA3 by Sex and ICU Status
Gene Expression
1.25
1.00
0.75
0.50
0.25
0.00
Female
Male
Sex
ICU Status
no
yes