Project 1_Submission1

Luwei Zeng

2024-07-28

Data is from this paper: https://pubmed.ncbi.nlm.nih.gov/33096026/

1. 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.
2. 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.

One gene:AIBG; One continuous covariate:ferritin(ng/ml); Two categorical covariates:sex, icu status

```
library("tidyverse")
```

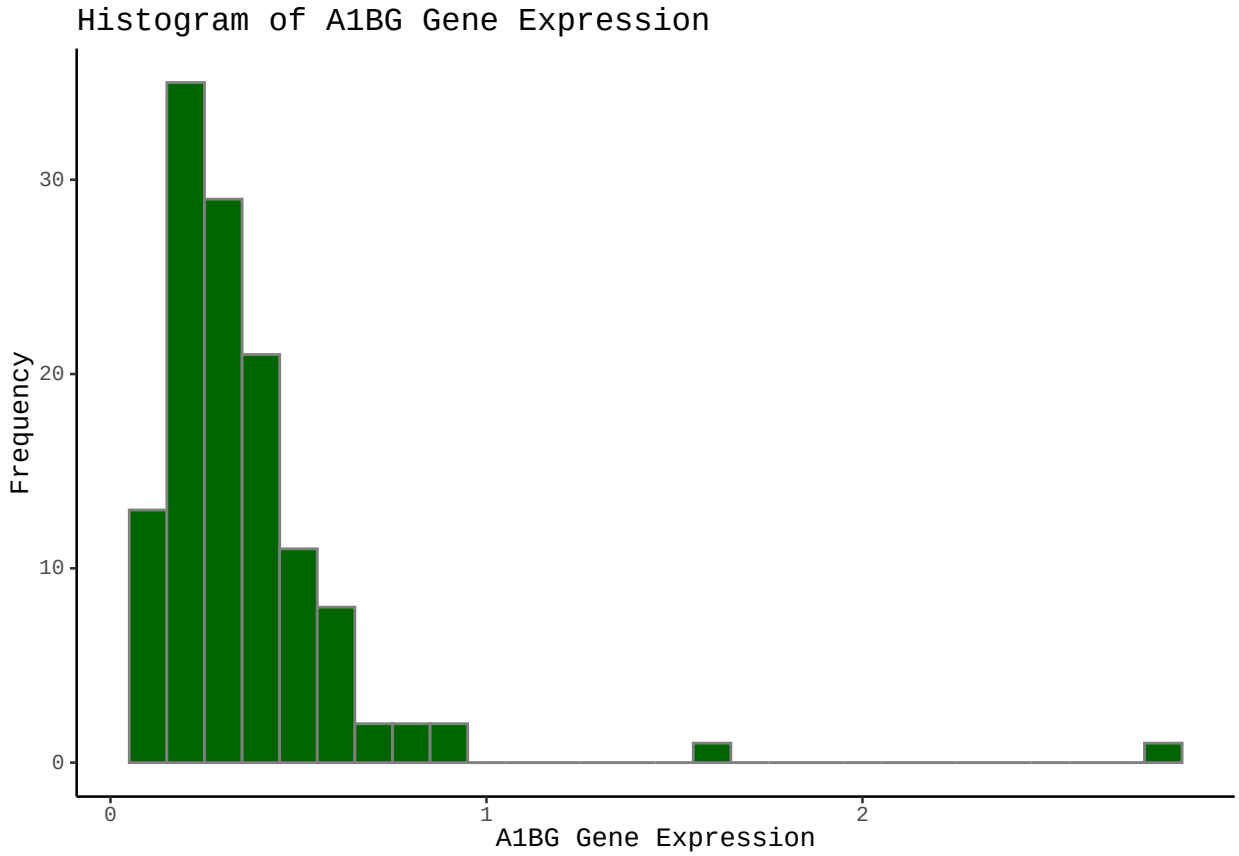
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ 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
```

3. Generate the following three plots using ggplot2 for your covariates of choice:

```
setwd("/Users/hekaiwei/Desktop/R class/project")
gene <- read.csv(file = "QBS103_GSE157103_genes.csv",row.names=1)
metadata <- read.csv(file = "QBS103_GSE157103_series_matrix.csv", row.names = 1)
#head(gene)
#head(metadata)

#convert to dataframe and switch column and row
A1BG_gene <- gene["A1BG", ]
A1BG_gene <- as.data.frame(t(A1BG_gene))
#head(A1BG_gene)

#combine A1BG gene expression and metadata
data_combined<-merge(A1BG_gene,metadata,by = "row.names")
colnames(data_combined)[1] <- "ID" #set the 1st column name as "ID"
#head(data_combined)
```

o Histogram for gene expression (5 pts)

```
#One gene:AIBG
#Define the overall plot with 'data_combined' as the data source and map 'A1BG' to the x-axis
ggplot(data_combined,aes(x = A1BG)) +
  #Add histogram data with specific bin width, fill color, and border color
  geom_histogram(binwidth=0.1,fill="darkgreen",color="grey50")+
  #Change labels for the title, x-axis, and y-axis
  labs(title="Histogram of A1BG Gene Expression",
     x="A1BG Gene Expression",
     y="Frequency")+
  # Define the theme as classic with 'Courier' font and base font size of 10
  theme_classic(base_family = 'Courier',base_size = 10)
```



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

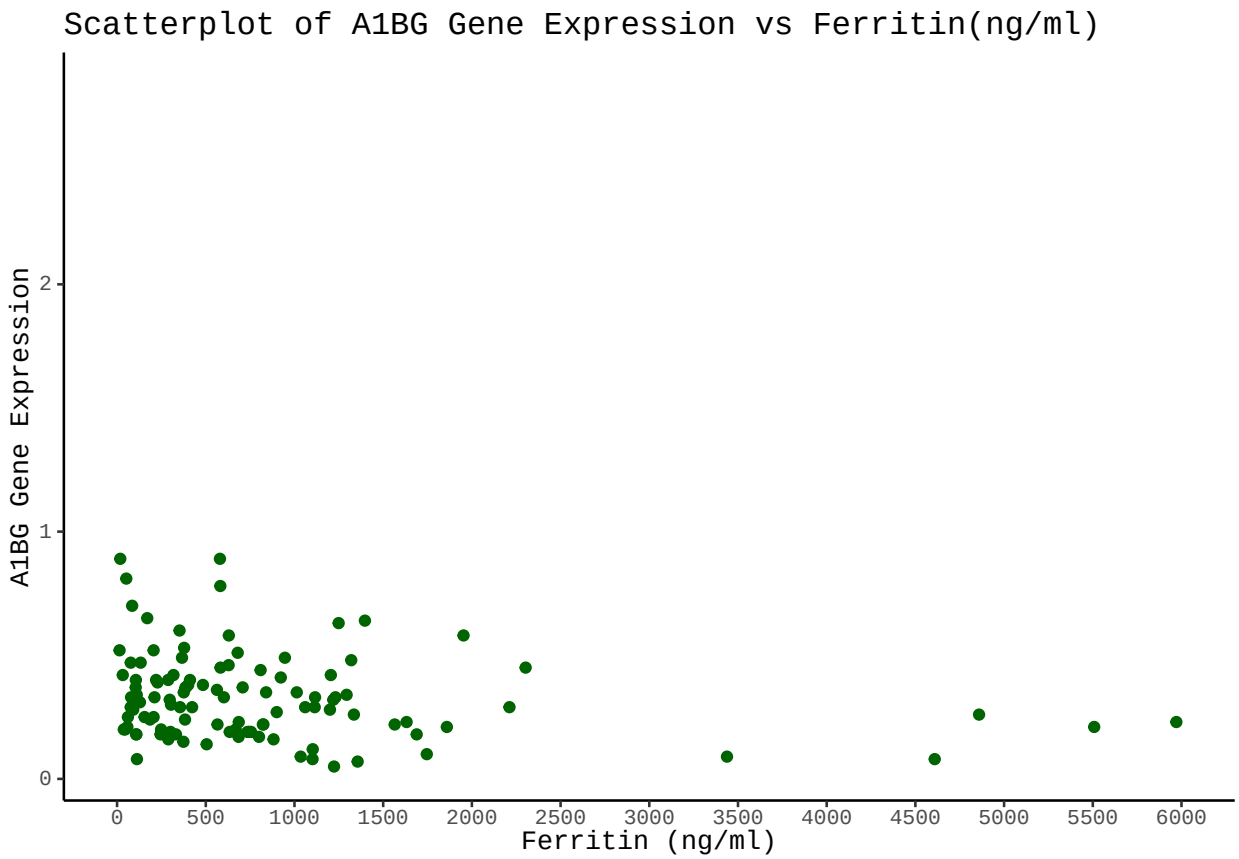
```
#One continuous covariate:ferritin(ng/ml)

#Ensure the column 'ferritin.ng.ml.' is of numeric type
data_combined$ferritin.ng.ml. <- as.numeric(data_combined$ferritin.ng.ml.)
```

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

```
#Define the overall plot with 'data_combined' as the data source and map 'ferritin.ng.ml.' to the x-axi
ggplot(data_combined, aes(x = ferritin.ng.ml., y = A1BG)) +
```

```
  #Add points to the plot, setting the color of the points to dark green
  geom_point(color="darkgreen") +
  #Change labels for the title, x-axis, and y-axis
  labs(title="Scatterplot of A1BG Gene Expression vs Ferritin(ng/ml)",
       x = "Ferritin (ng/ml)",
       y = "A1BG Gene Expression") +
  #Define the scale for the x-axis, setting limits from 0 to 6000 and breaks at intervals of 500
  scale_x_continuous(limits = c(0, 6000), breaks = seq(0, 6000, by = 500)) +
  # Define the theme as classic with 'Courier' font and base font size of 10
  theme_classic(base_family = 'Courier',base_size = 10)
```

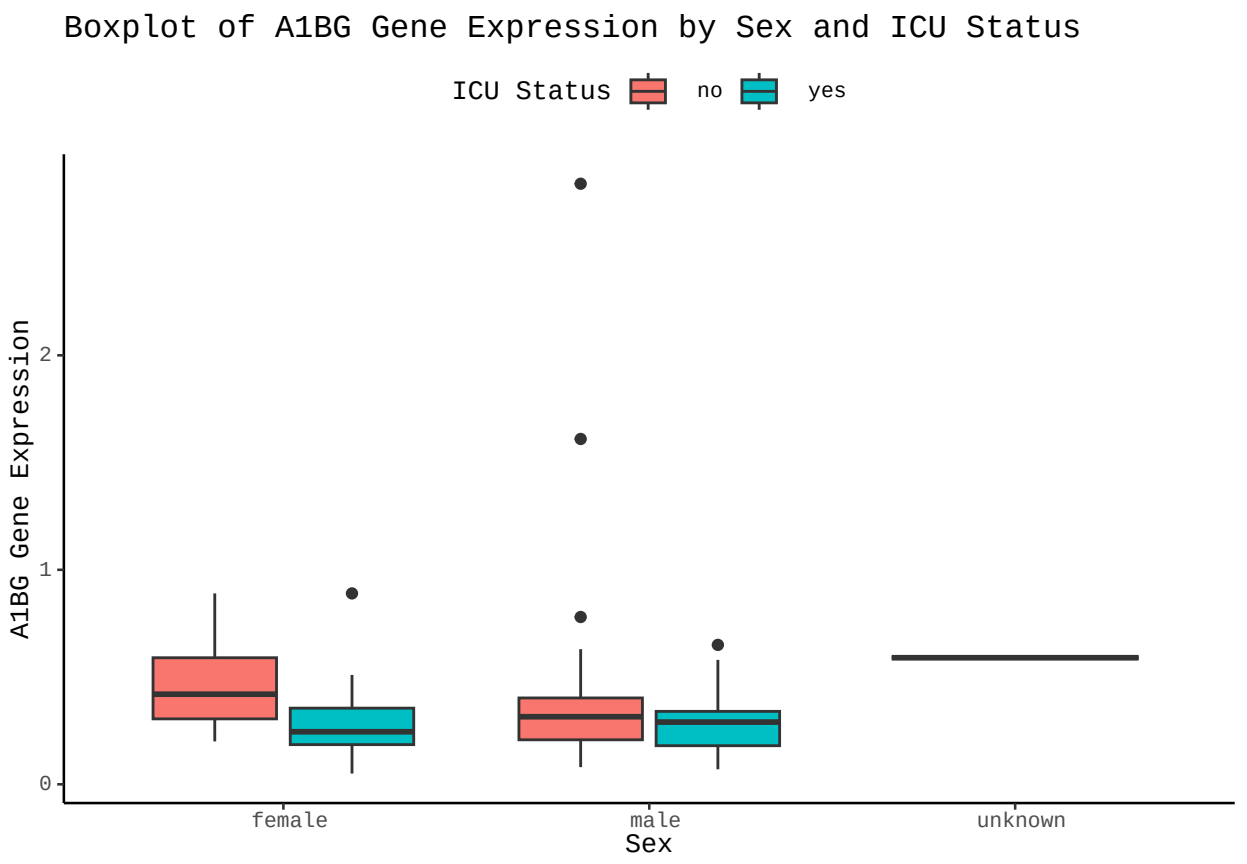
```
## Warning: Removed 16 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



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

```
#Two categorical covariates sex, icu status
#Define the overall plot with 'data_combined' as the data source and map 'sex' to the x-axis, 'A1BG' to
ggplot(data_combined, aes(x = sex, y = A1BG, fill = icu_status)) +
  # Add boxplot
  geom_boxplot()+
  #Change labels for the title, x-axis, y-axis, and fill legend
  labs(title = "Boxplot of A1BG Gene Expression by Sex and ICU Status",
       x = "Sex",
       y = "A1BG Gene Expression",
```

```
      fill = "ICU Status") +
  #Define the theme as classic with 'Courier' font and base font size of 10
  theme_classic(base_family = 'Courier',base_size = 10)+
  # Customize the theme to position the legend at the top of the plot
  theme(legend.position = 'top')
```



4. 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)

5. Submit your clearly commented code and generated plots as a knitted R markdown file.