Final submission

Ethan Zhang

2024-08-25

```
new_gene <-
  read.csv(
    "/Users/zhyihan/Documents/Dartmouth Life/courses/QBS 103-Data Science/final project/QBS103_GSE157103
  )
ser_m <-
  read.csv(
    "/Users/zhyihan/Documents/Dartmouth Life/courses/QBS 103-Data Science/final project/QBS103_GSE157103
  )

new_gene_long <-
  new_gene %>% gather(
    key = participant_id,
    value = expression,
    COVID_01_39y_male_NonICU:NONCOVID_26_36y_male_ICU
  )

new_data <- merge(ser_m, new_gene_long, by = "participant_id")
new_data <- new_data %>% filter(X == 'ABCA1')
```

Generate a table formatted in LaTeX of summary statistics for all the covariates you looked at and 2 additional continuous (3 total) and 1 additional categorical variable (3 total). (5 pts) o Stratifying by one of your categorical variables o Tables should report n (%) for categorical variables o Tables should report mean (sd) or median [IQR] for continuous variables

```
options(warn = -1)
new_ser_m<- ser_m%>% select("age","ferritin.ng.ml.","crp.mg.l.", "sex", "icu_status", "charlson_score")
new_ser_m$age <- as.numeric(new_ser_m$age)
new_ser_m$ferritin.ng.ml. <- as.numeric(new_ser_m$ferritin.ng.ml.)
new_ser_m$crp.mg.l. <- as.numeric(new_ser_m$crp.mg.l.)

char <- data.frame(matrix(ncol = 2, nrow = 0),stringsAsFactors = FALSE)

age <- c("Age",paste0(round(mean(new_ser_m[which(new_ser_m$icu_status == ' yes'),]$age,na.rm = TRUE),2)

fer<-c("Ferritin (ng/mL)",paste0(round(mean(new_ser_m[which(new_ser_m$icu_status == ' yes'),]$ferritin.n

crp <- c("CRP (mg/L)", paste0(round(mean(new_ser_m[which(new_ser_m$icu_status == ' yes'),]$crp.mg.l.,na


Sex <- c("Sex", "","")
char <- rbind(char,age,fer,crp,Sex, stringsAsFactors = FALSE)
```

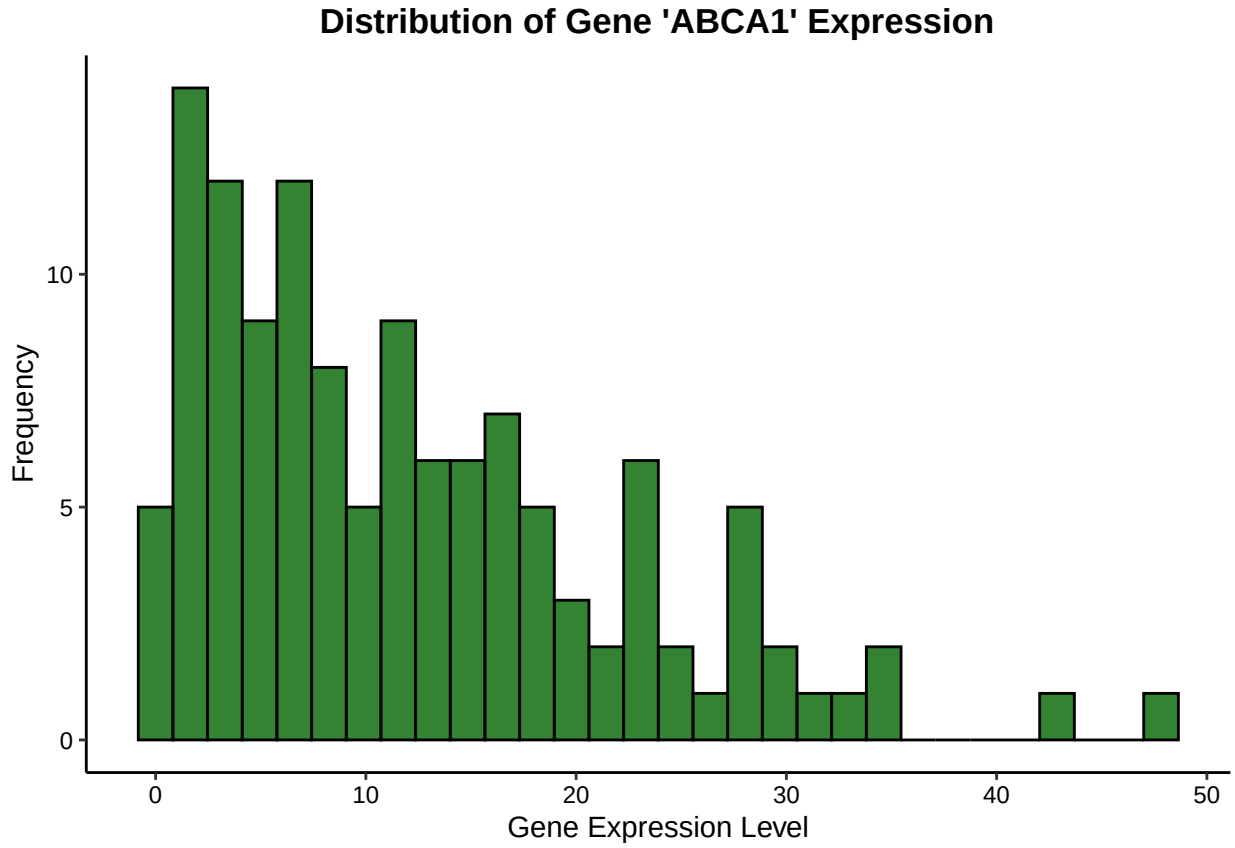
ICU Status Stratification (mean (sd))		
	In ICU	Not In ICU
Age	63.45 (14)	58.67 (17.82)
Ferritin (ng/mL)	935.32 (1019.02)	715.75 (1067.55)
CRP (mg/L)	149.57 (105.54)	109.4 (94.38)
Sex		
male	41 (62.12%)	33 (55%)
female	24 (36.36%)	27 (45%)
unknown	1 (1.52%)	0 (0%)
Charlson Score		
0	1 (1.52%)	10 (16.67%)
1	10 (15.15%)	9 (15%)
2	14 (21.21%)	11 (18.33%)
3	13 (19.7%)	3 (5%)
4	6 (9.09%)	9 (15%)
5	6 (9.09%)	5 (8.33%)
6	2 (3.03%)	6 (10%)
7	8 (12.12%)	4 (6.67%)
8	3 (4.55%)	3 (5%)
9	2 (3.03%)	0 (0%)
11	1 (1.52%)	0 (0%)

```
cate <- new_ser_m$sex
for(i in unique(cate)){
  col <- c(i,paste0(round(nrow(new_ser_m[which(new_ser_m$icu_status == ' yes' & cate == i),]),2),' (',r
  char <- rbind(char,col,stringsAsFactors = FALSE)
}

char <- rbind(char,c("Charlson Score","",""),stringsAsFactors = FALSE)

cate <- new_ser_m$charlson_score
for(i in sort(unique(cate))){
  col <- c(i,paste0(round(nrow(new_ser_m[which(new_ser_m$icu_status == ' yes' & cate == i),]),2),' (',r
  char <- rbind(char,col,stringsAsFactors = FALSE)
}

names(char) <- c("", "In ICU", "Not In ICU")

tab <- char %>% kbl(format = "latex") %>% kable_classic(c('striped', 'condensed'), full_width = F, font
  add_header_above(c("ICU Status Stratification (mean (sd))" = 3),color="white", background="#7B7B7B")%

tab
```

```
#define a theme for all plots
New_theme <- theme(
  panel.border = element_blank(),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
```

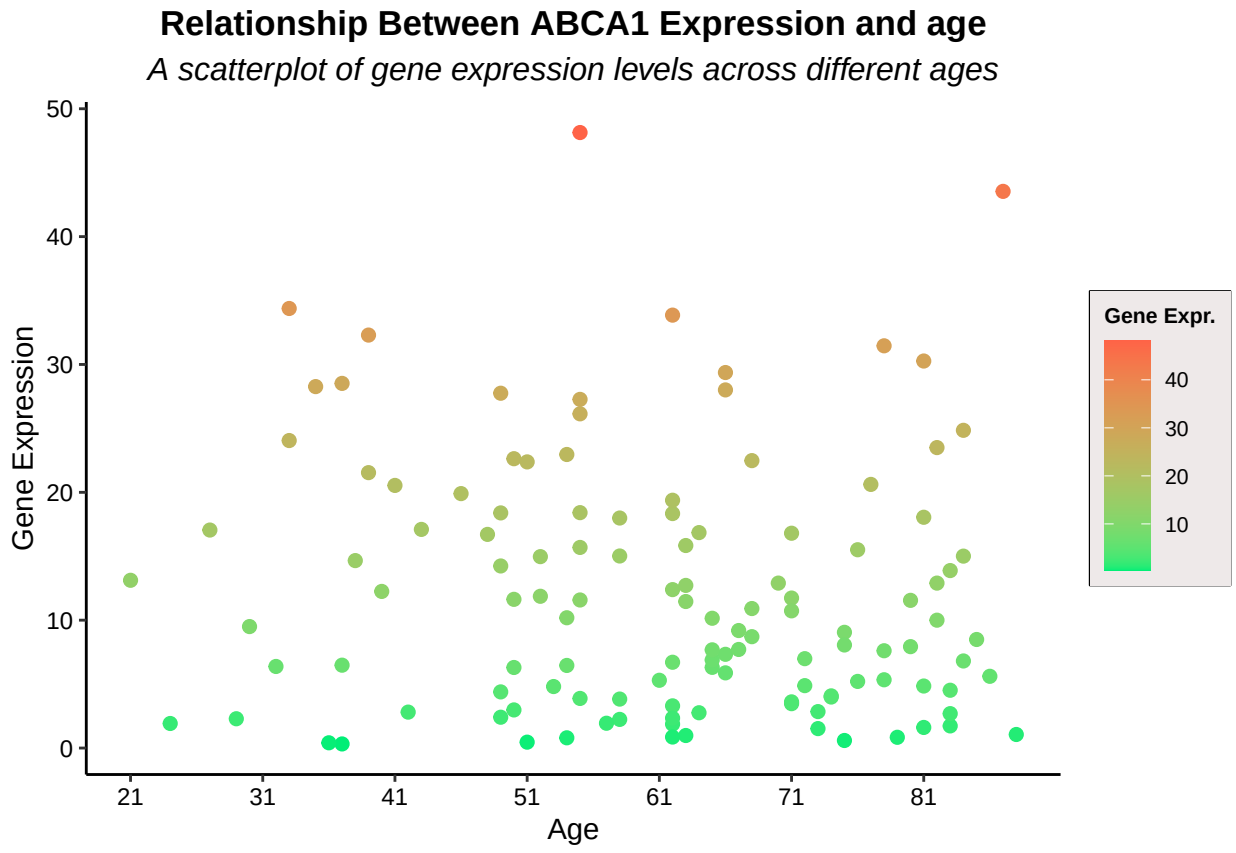
```
  # Set plot background
  plot.background = element_rect(fill = "white"),
  panel.background = element_blank(),
  legend.background = element_rect(fill = 'snow2'),
  legend.text = element_text(color = "black", size = 8),
  legend.title = element_text(color = "black", face = "bold", size = 8),
  legend.key = element_rect(fill = "snow2",color = "snow2"),
  legend.box.background  = element_rect(color = "black"),
  ##make the title center
  plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
  plot.subtitle = element_text(hjust = 0.5, size = 12, face = "italic"),
  title = element_text(color = "black"),
  axis.line = element_line(color = "black"),
  axis.text = element_text(color = "black"),
  legend.position = 'right'
)
```

```
####Histogram for gene expression
bar_char <- ggplot(new_data, aes(x=expression)) + geom_histogram(bins =30, fill = "darkgreen", color =
       x = "Gene Expression Level",
       y = "Frequency") + New_theme

bar_char
```



```
#ggsave("/Users/zhyihan/Documents/Dartmouth Life/courses/QBS 103-Data Science/final project/final_bar_c
```

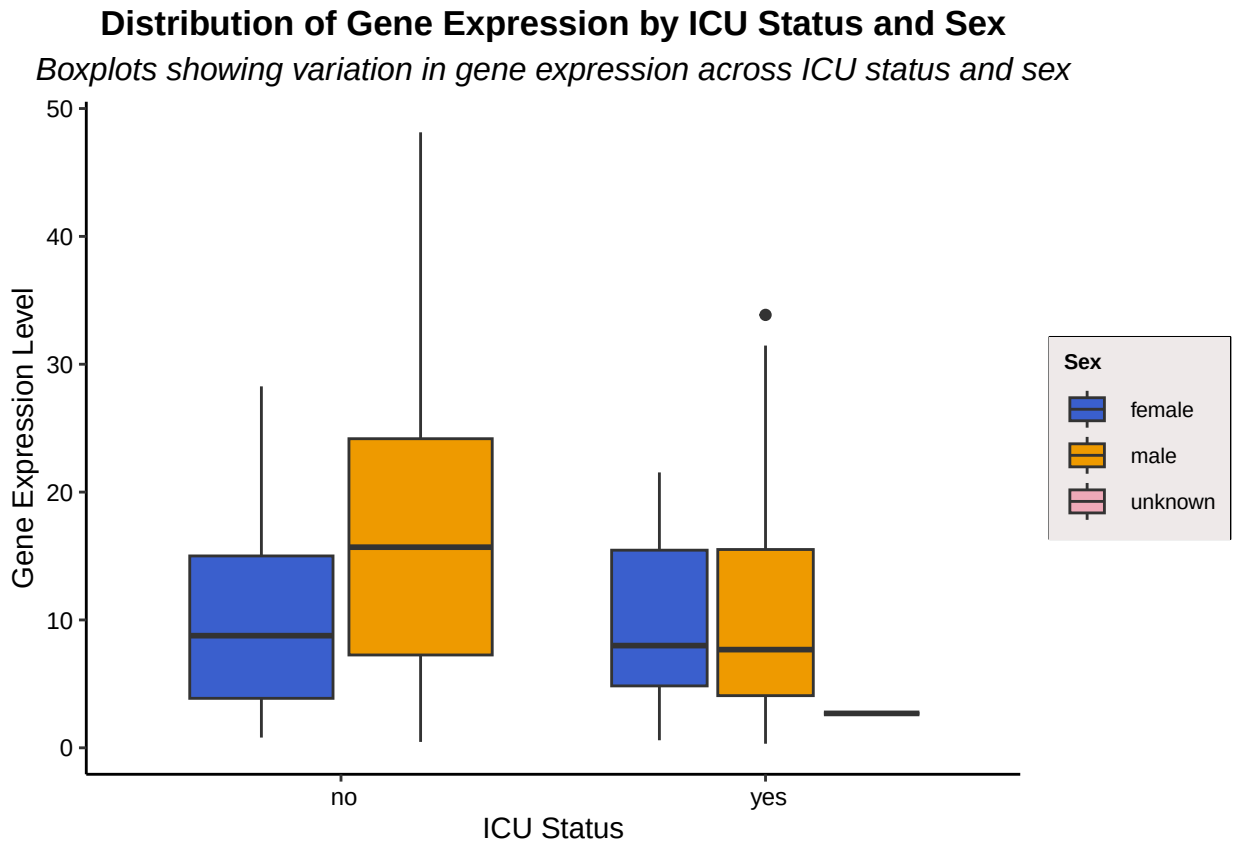
```
####Scatterplot for gene expression and continuous covariate
##I make the color fade from green to tomato
new_data$age[!grepl("^[0-9]+$", new_data$age)] <- NA
new_data<-new_data%>%drop_na()
new_data$age <- as.character(new_data$age)
new_data$age <- as.numeric(new_data$age)
scatter <- ggplot(new_data,
       aes(x = age, y = expression, color = expression)) + geom_point(size = 2) + labs(
         title = "Relationship Between ABCA1 Expression and age",
         subtitle = "A scatterplot of gene expression levels across different ages",
         y = 'Gene Expression',
         x = 'Age',
         color = 'Gene Expr.'
       ) + scale_color_gradient(low = "springgreen2", high = "tomato1") + New_theme +scale_x_continuous
scatter
```




```
#ggsave("/Users/zhyihan/Documents/Dartmouth Life/courses/QBS 103-Data Science/final project/final_scatt
```

```
####Boxplot of gene expression separated by both categorical covariates
box_plot<- ggplot(new_data,aes(x = icu_status ,y = expression,fill = sex)) +geom_boxplot()+ New_theme +
    title = "Distribution of Gene Expression by ICU Status and Sex",
    subtitle = "Boxplots showing variation in gene expression across ICU status and sex",
    x = "ICU Status",
```

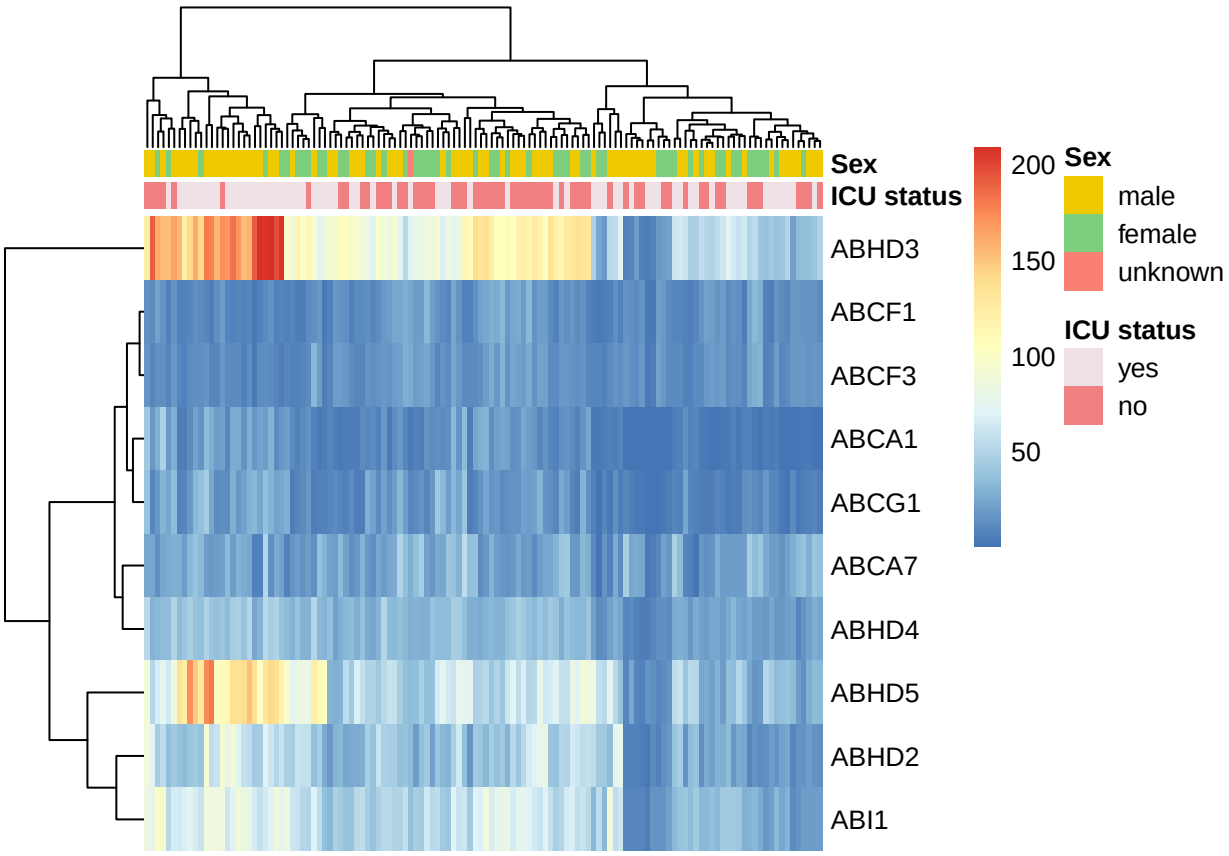
```
    y = "Gene Expression Level",
    fill = "Sex"
  ) +scale_fill_manual(values = c('royalblue3', 'orange2','pink2'))
print(box_plot)
```




```
#ggsave("/Users/zhyihan/Documents/Dartmouth Life/courses/QBS 103-Data Science/final project/final_box_p
```

```
# Generate heatmap without clustering
test_data<-new_gene %>% filter(X == 'ABI1'|X =='ABHD2'|X =='ABHD3'|X =='ABHD4'|X =='ABHD5'|X =='ABCG1'|X
rownames(test_data) <- test_data$X
test_data<-test_data[,-1]


annotation <- data.frame(
  `ICU status` = factor(ser_m$icu_status),
  Sex = factor(ser_m$sex),
  check.names = FALSE
)
rownames(annotation) <- colnames(test_data)

#fill_colors <- colorRampPalette(c('pink2','orange2','royalblue3'))(10)

annot_color <- list(
  Sex = c(" male" = 'gold2', " female" = 'palegreen3', " unknown" = "salmon"),
  `ICU status` = c(" yes" = "lavenderblush2", " no" = "lightcoral")
```

```
)


# Use pheatmap with the correct color parameter
heat_map <- pheatmap(
  test_data,
  clustering_distance_cols = 'euclidean',
  clustering_distance_rows = 'euclidean',
  fontsize_col  = 2,
  annotation_col = annotation,
  show_colnames = FALSE,
  #color = fill_colors,
  annotation_colors = annot_color
)
```




```
#ggsave("/Users/zhyihan/Documents/Dartmouth Life/courses/QBS 103-Data Science/final project/final_heatm
```

```
###delete unknown sex
new_data3 <- new_data[which(new_data$sex != ' unknown'),]

# Create violin plot
violin_plot <- ggplot(new_data3, aes(x = factor(sex), y = expression, fill = factor(sex))) +
  geom_violin(trim = FALSE) +
  labs(
    title = "Comparison of Gene Expression Levels by Sex",
```

```
    x = "Sex",
    y = "Gene Expression",
    fill = "Sex"
  ) +
  New_theme+
  scale_fill_manual(values = c("skyblue", "pink"))

# Display the plot
print(violin_plot)
```


```
#ggsave("/Users/zhyihan/Documents/Dartmouth Life/courses/QBS 103-Data Science/final project/final_violi
```