Final Project

2025-07-14

```r
# Read gene data from file
genes<- read.csv("QBS103_GSE157103_genes.csv", header = TRUE)
# Read participant data from file
matrix<- read.csv("QBS103_GSE157103_series_matrix-1.csv", header = TRUE)
```

```r
# Transpose the gene data to make rows participant data
genes_trans <- as.data.frame(t(genes))
colnames(genes_trans) <- as.character(genes_trans[1,])
genes_trans <- genes_trans[-1,]

# Merge gene data with participant data
genes_trans$participant_id <- rownames(genes_trans)
rownames(matrix)<- matrix$participant_id
merged_df <- merge(genes_trans, matrix , by = "row.names")
colnames(merged_df)[1] <- "participant_id"
```

```r
### Selecting a gene and 3 covariates
selected_cov <- c("age", "sex", "disease_status", "ABHD14B")
subset_merged <- merged_df[,selected_cov]
subset_merged$age <- as.numeric(subset_merged$age)
```

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

```r
subset_merged$ABHD14B <- as.numeric(subset_merged$ABHD14B)
subset_merged <- na.omit(subset_merged)
```

```r
library(ggplot2)
subset_merged$ABHD14B <- as.numeric(subset_merged$ABHD14B)

ggplot(subset_merged, aes(x = ABHD14B)) +
  geom_histogram(binwidth = 2.0, fill = "steelblue", color = "black") +
  labs(title = "Gene Expression of ABHD14B", x = "ABHD14B", y = "Count")
```
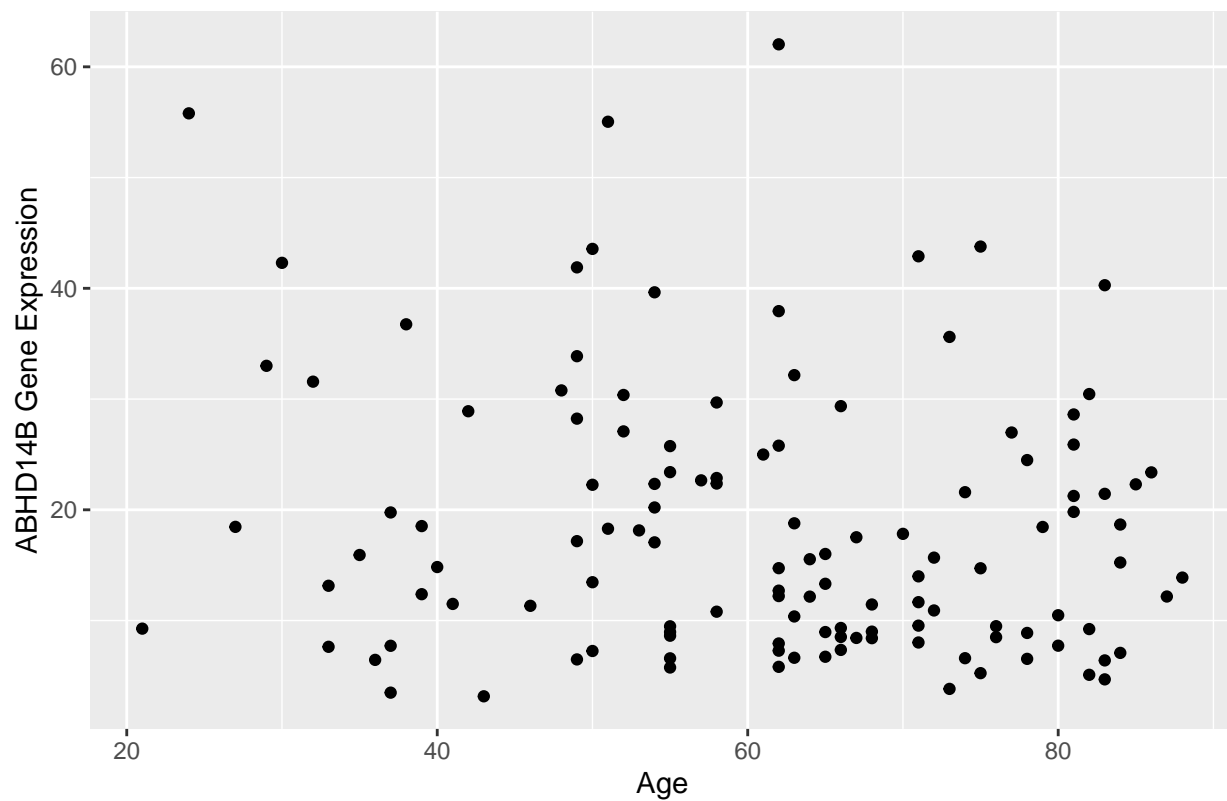
Gene Expression of ABHD14B
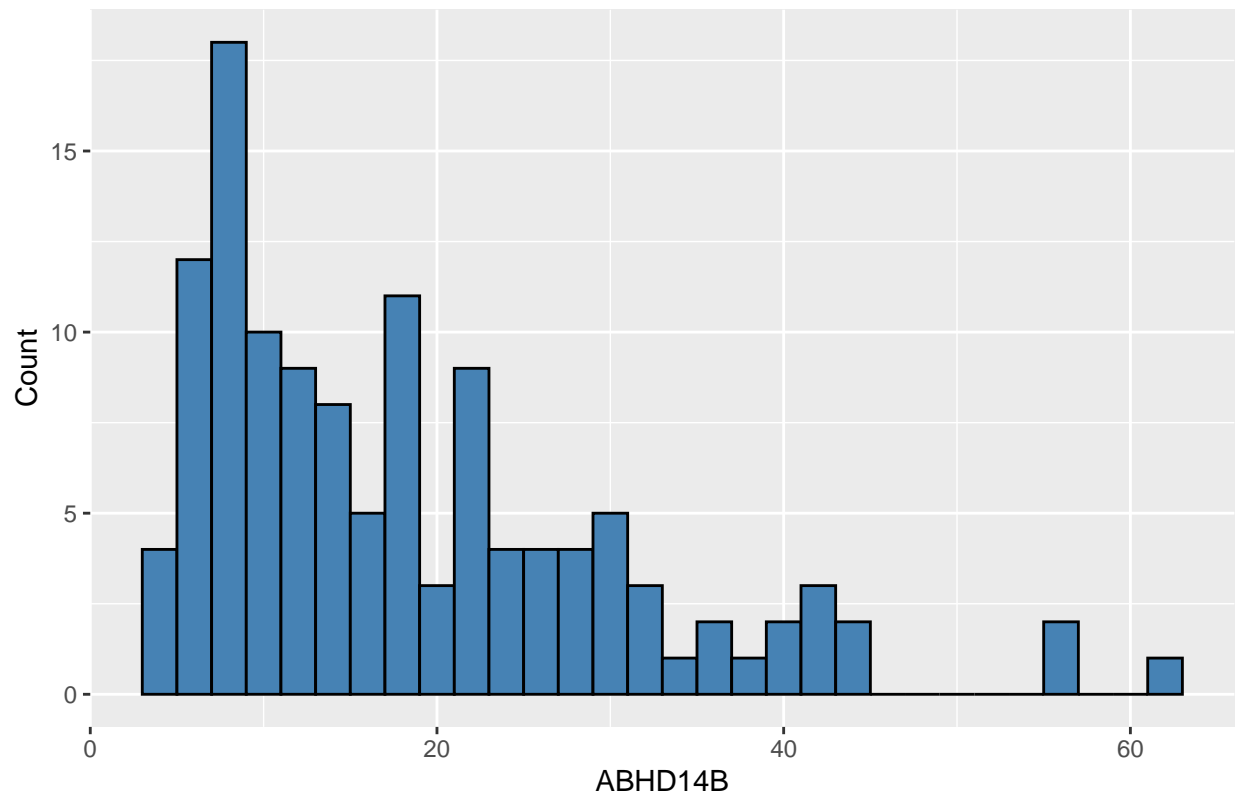

```r
ggplot(subset_merged, aes(x = age, y = ABHD14B)) +
  geom_point() +
  labs(title = "Scatterplot b/w Gene Expression and Age", y = "ABHD14B Gene Expression", x = "Age")
```

Scatterplot b/w Gene Expression and Age


```r
subset_merged = subset_merged[subset_merged$sex != " unknown", ]

ggplot(subset_merged, aes(x = sex, y = ABHD14B, fill = disease_status)) +
  geom_boxplot() +
  labs(title = "ABHD14B Gene Expression by Sex and Disease Status",
       x = "Sex", y = "ABHD14B Gene Expression")
```

ABHD14B Gene Expression by Sex and Disease Status