Project: Visualizing Gene Expression in Breast Cancer using ggplot2

Objective:

- Analyze and visualize gene expression patterns between breast cancer tissues and normal tissues using the ggplot2 package in R
- The following plots have to be included-

1. Boxplot (per gene)
2. Violin Plot (per gene).

GEO Dataset used - GSE15852 (Expression data from human breast tumors and their paired normal tissues)

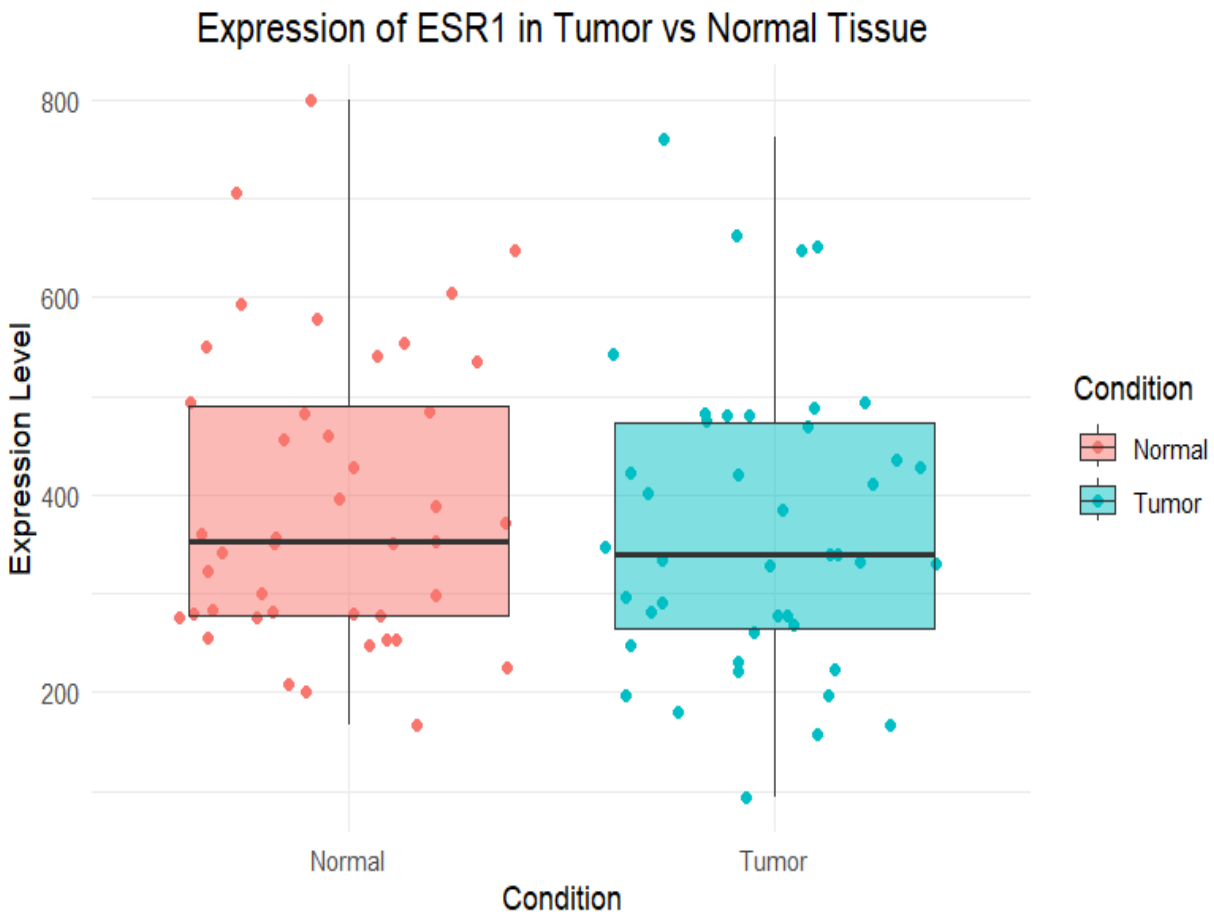
Steps before plotting

```
# Load necessary libraries
library(GEOquery) # downloads GEO Dataset
library(ggplot2) # used for plotting graph
library(dplyr) # used for data manipulation
library(tidyr) # used to keep the data clean
# Download dataset from GEO
gse <- getGEO("GSE15852", GSEMatrix = TRUE)
# Extract expression matrix (genes x samples)
expr_matrix <- exprs(gse[[1]])
# Extract sample metadata
metadata <- pData(gse[[1]])
# Create a Condition column based on 'title' field
metadata$Condition <- ifelse(grepl("normal", metadata$title, ignore.case = TRUE), "Normal", "Tumor“)
```

Continued..

```
# Convert matrix to data frame and add gene names

expr_df <- as.data.frame(expr_matrix)

expr_df$Gene <- rownames(expr_df)

# Convert to long (tidy) format: Gene | Sample | Expression

long_expr <- pivot_longer(expr_df, cols = -Gene,  names_to = "Sample", values_to = "Expression")

# Add Tumor/Normal condition to expression

 datalong_expr <- left_join(long_expr, metadata %>% select(geo_accession, Condition),by = c("Sample" = "geo_accession"))

# Add Tumor/Normal condition to expression

 datalong_expr <- left_join(long_expr, metadata %>% select(geo_accession, Condition),  by = c("Sample" =
"geo_accession"))

# Choose a gene to visualizegene_of_interest <- "1861_at“,  Filter data to only this gene

gene_data <- long_expr %>%  filter(Gene == gene_of_interest)head(unique(long_expr$Gene), 20) # to check different
genesnames (gene_data)
```

Box plot (ESR1 gene)



- Code-

```
a<ggplot(data=gene_data,aes(x=Condition,y=Expression,fill=Condition))

a+geom_jitter(aes(colour=Condition))+geom_boxplot(alpha=0.5)+

xlab("Condition")+  ylab("Expression Level")+

 ggtitle("Expression of ESR1 in Tumor vs Normal Tissue")+

 theme_minimal()+

theme(plot.title = element_text(hjust = 0.5))
```

Interpretation of Box Plot

Code-
t.test(Expression ~ Condition, data = gene_data)
table(gene_data$Condition)

t-test Results:
t = 0.80276
p-value = 0.4244
mean in Normal = 392.19
mean in Tumor = 366.42
95% CI = [-38.08, 89.63]

The boxplot compares **ESR1** expression between **Tumor** and **Normal** breast tissue samples (n = 43 each) from the GSE15852 dataset.

•The **Tumor group** shows a slightly **wider and thinner interquartile range**, indicating more variability in expression levels.

•The **Normal group** has more **outliers** beyond the whiskers, suggesting some samples with unusually high or low ESR1 expression.

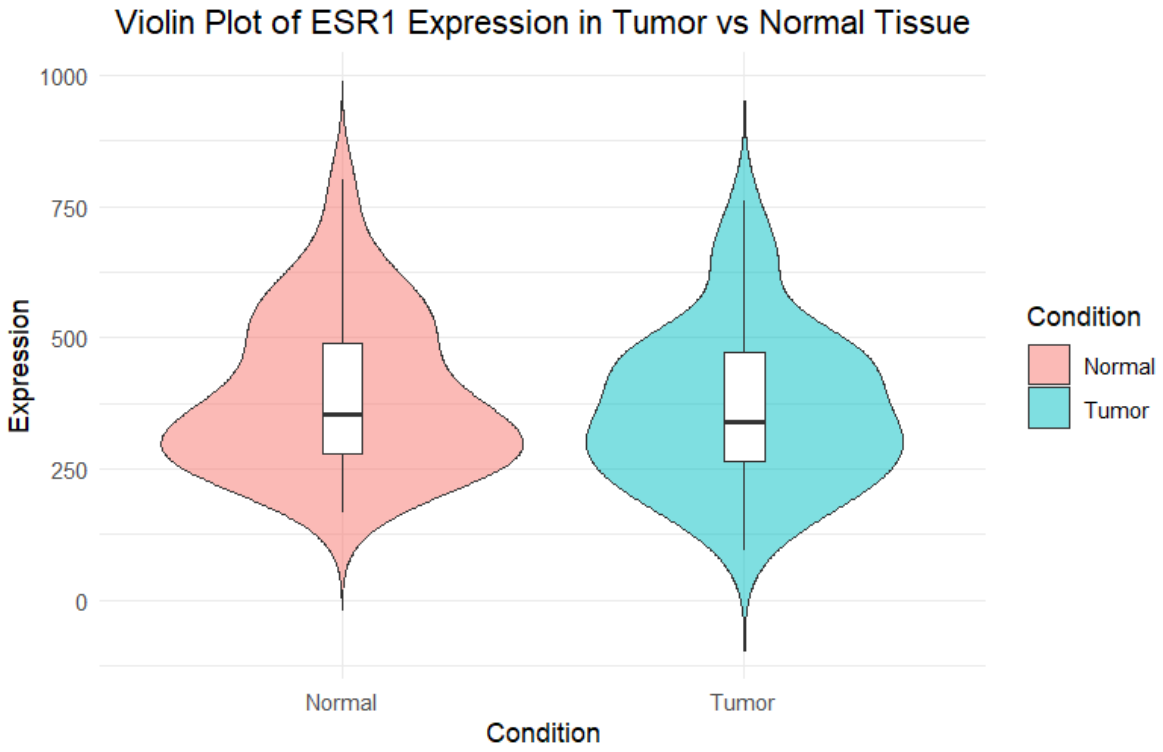
•The median expression is **slightly higher in Normal samples** compared to Tumor, but this difference is **not statistically significant** (Welch t-test p = 0.424).

•The 95% confidence interval for the mean difference (-38.08 to 89.63) includes zero, supporting the lack of significant difference.

•Biologically, ESR1 is often upregulated in estrogen receptor–positive breast tumors, but this dataset likely includes multiple tumor subtypes, which may explain the absence of a clear difference.

Overall, this analysis suggests **no strong evidence of differential ESR1 expression** between tumor and normal tissues in this cohort.

Violin plot



Code-

```
a<ggplot(data=gene_data,aes(x=Condition,y=Expression,fill=Conditi
on))

a+geom_violin(trim = FALSE,alpha=0.5)+
geom_boxplot(width=0.1,fill="white")+

ggtitle("Violin Plot of ESR1 Expression in Tumor vs Normal Tissue")
+

 theme_minimal() +

theme(plot.title = element_text(hjust = 0.5))
```

Interpretation-

•ESR1 expression shows **similar distribution** in Tumor and Normal tissue
•**Median expression** is slightly higher in Normal, but not significantly.
•Normal samples show **slightly more variability**.
•Pattern supports the **t-test result (p = 0.424)** — no significant difference.

Bar plot