Test

Data Source

This study will be using the data from the University of Texas Health Science Center at Houston, which is used for research on how miRNA-138 suppresses glioblastoma proliferation through downregulation. The researchers have already cleaned the data, and the data given to us is already cleaned. However, we are uncertain about the data collection process or the data cleaning that they conducted.

Data Dimensions

The cleaned dataset consists of 627 different miRNAs (rows) and data from 4 healthy individuals and 9 glioblastoma patients (columns). The data has two types of variables: numerical and categorical. The name of each miRNA is coded as text, which is considered a categorical variable. The values that represent expression levels for the control group and the glioblastoma groups are all integers.

Validated Data Cleaning

We have checked that the data does not contain any duplicate miRNA; in other words, each row represents a unique miRNA. Moreover, we also verified that the values for the control and glioblastoma columns are all non-negative, and there are no missing values in this dataset. In this EDA section, there is no distinct outlier since the expression levels of a miRNA vary a lot. Hence, no data was removed, and we did not transform the data.

```
# # Check if col1 has duplicates
# col1_duplicates <- any(duplicated(df$col1))
# if (col1_duplicates) {
#   print("Column 1 has duplicates.")
# } else {
#   print("Column 1 has no duplicates.")
```

```
# }
#
# # Check for negative numbers or NAs in the rest of the columns
# for (col in colnames(df)[-1]) {  # Exclude the first column
#   neg_values <- any(df[[col]] < 0, na.rm = TRUE)
#   na_values <- any(is.na(df[[col]]))
#
#   if (neg_values) {
#     print(paste("Column", col, "contains negative numbers."))
#   }
#
#   if (na_values) {
#     print(paste("Column", col, "contains NA values."))
#   }
# }
```