STA490 EDA

Descriptive Statistics

```
# 1. Summary Statistics for Numerical Variables
summary_stats <- data %>%
  select(-miRNA_ID) %>%
  summarise(
    mean = colMeans(.),
    median = apply(., 2, median),
    sd = apply(., 2, sd),
    min = apply(., 2, min),
    max = apply(., 2, max)
  ) %>%
  t() %>%
  as.data.frame()

# Add row names as a column
summary_stats <- cbind(Statistic = rownames(summary_stats), summary_stats)
rownames(summary_stats) <- NULL

# Rename columns for clarity
colnames(summary_stats) <- c("Statistic", colnames(data)[-1])

# Split the data into two tables
control_table <- summary_stats %>% select(Statistic, Ctrl1, Ctrl2, Ctrl3, Ctrl4, GBM1, GBM2,
gbm_table <- summary_stats %>% select(Statistic, GBM4, GBM5, GBM6, GBM7, GBM8, GBM9)

# Generate the first table with the caption
control_table %>%
  kbl(caption = "Summary Statistics for Numerical Variables") %>%
  kable_styling(full_width = FALSE, latex_options = "scale_down")
```

Table 1: Summary Statistics for Numerical Variables

Statistic	Ctrl1	Ctrl2	Ctrl3	Ctrl4	GBM1	GBM2	GBM3
mean	931.0861	1060.550	1079.829	821.5614	769.1244	657.4434	1208.571
median	16.0000	20.000	20.000	19.0000	16.0000	20.0000	21.000
sd	7922.3853	7388.496	6798.288	5451.7693	6145.5172	4570.3958	9243.272
min	1.0000	2.000	0.000	2.0000	3.0000	1.0000	2.000
max	177267.0000	135184.000	96358.000	95868.0000	130396.0000	94916.0000	165684.000

Statistic	GBM4	GBM5	GBM6	GBM7	GBM8	GBM9
mean	1398.136	986.8692	782.7337	960.3301	951.2249	1133.475
median	19.000	20.0000	18.0000	17.0000	19.0000	17.000
sd	19286.050	8626.0398	5643.0384	6845.7946	5847.1270	8237.234
min	0.000	2.0000	3.0000	1.0000	2.0000	0.000
max	474457.000	185424.0000	92773.0000	106511.0000	86023.0000	103717.000

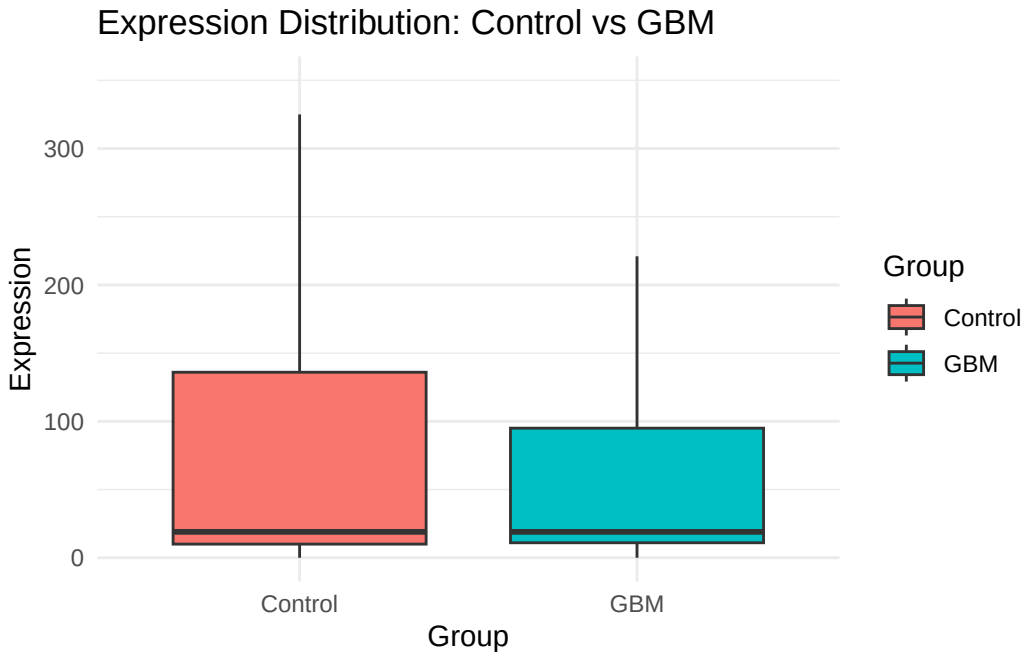
```
# Generate the second table without a caption
gbm_table %>%
  kbl() %>%
  kable_styling(full_width = FALSE, latex_options = "scale_down")
```

As the descriptive statistics shown, significant variability in miRNA expression levels between control and GBM samples is revealed with GBM exhibiting greater heterogeneity, as indicated by higher standard deviations (e.g., GBM4 with ~19,286). The median expression values are much lower than the means across all samples, suggesting a right-skewed distribution due to a few highly expressed miRNAs. The maximum expression levels in GBM samples (e.g., GBM4 at ~474,457) underscore the extreme upregulation of certain miRNAs, consistent with the known heterogeneity and adaptability of glioblastoma cells. These findings support the potential critical role of miRNAs in regulating glioblastoma stem cells (GSCs), influencing tumor growth, survival, and therapy resistance. This variability highlights miRNAs as potential therapeutic targets, warranting further exploration to identify specific miRNAs that could disrupt GSC functions and improve GBM treatment outcomes.

```
# 2. Distribution of Control vs GBM Values
control_columns <- grep("Ctrl", colnames(data), value = TRUE)
gbm_columns <- grep("GBM", colnames(data), value = TRUE)

data_distribution <- data %>%
  select(all_of(control_columns), all_of(gbm_columns)) %>%
  reshape2::melt() %>%
  mutate(Group = ifelse(grepl("Ctrl", variable), "Control", "GBM"))
```

```
No id variables; using all as measure variables
```

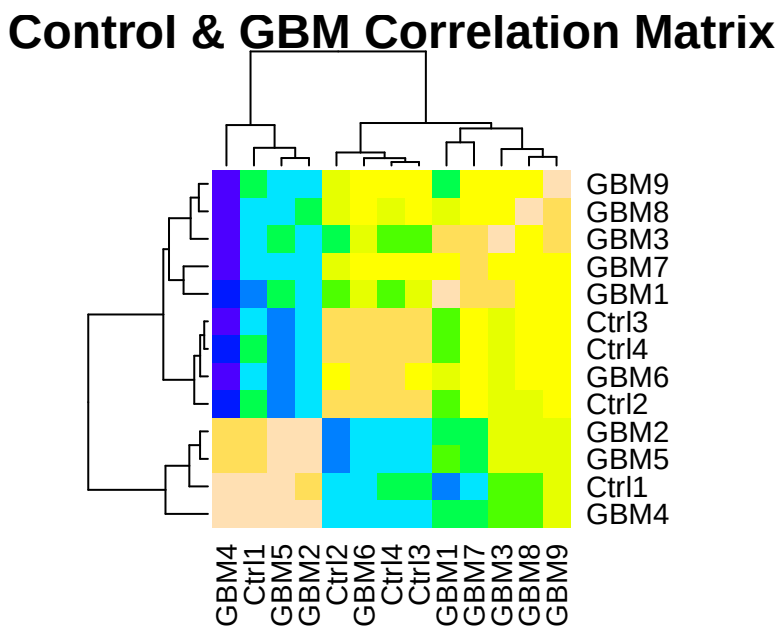
```r
# Plot the distribution without outliers and limit y-axis
ggplot(data_distribution, aes(x = Group, y = value, fill = Group)) +
  geom_boxplot(outlier.shape = NA) +
  theme_minimal() +
  labs(title = "Expression Distribution: Control vs GBM",
       x = "Group", y = "Expression") +
  coord_cartesian(ylim = c(0, 350)) +
  stat_summary(fun = mean, geom = "point", shape = 20, size = 3,
               color = "red", fill = "red")
```



In this analysis, the data was divided into two groups: **Control** and **GBM**, by selecting the columns labeled "Ctrl" and "GBM" respectively. The expression values were reshaped into a long format to facilitate comparison. A boxplot was generated to visualize the distribution of expression levels for each group, excluding outliers for a clearer representation. The y-axis range was restricted to 0–350 to focus on the primary distribution and exclude extreme values. Red points were added to represent the mean expression for each group.

The boxplot reveals that **Control samples** generally exhibit higher variability in expression levels compared to GBM samples, as evidenced by the wider interquartile range (IQR) for the Control group. The median expression levels appear similar between the two groups, but the GBM group shows a slightly tighter distribution. These findings suggest that GBM cells have less variation in miRNA expression within the observed range, potentially reflecting their more homogeneous cellular behavior compared to the heterogeneous Control group.

```r
# 3. Correlation Analysis between Numerical Variables
correlation_matrix <- data %>%
  select(-miRNA_ID) %>%
  cor()

# Visualize the Correlation Matrix
heatmap(as.matrix(correlation_matrix),
        main = "Control & GBM Correlation Matrix",
        col = topo.colors(10))
```



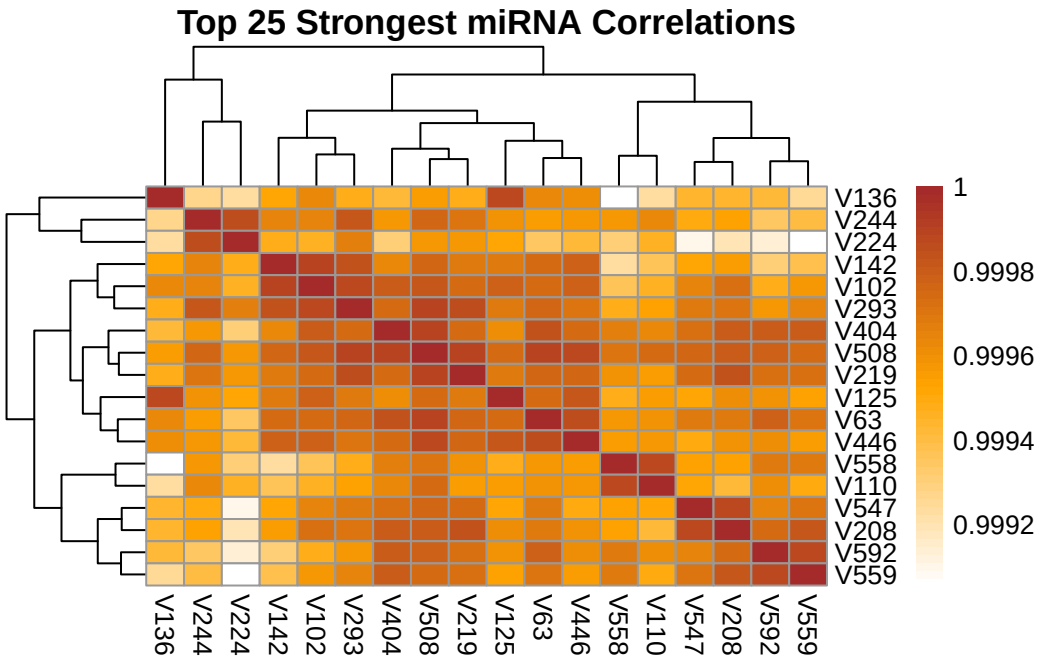
The correlation matrix reveals distinct clustering of control and GBM samples, with each group showing strong intra-group correlations, as indicated by the prevalence of warmer colors (yellow) within their respective clusters. The weaker inter-group correlations, represented by cooler colors (green/blue), highlight the distinct miRNA expression profiles between control and GBM samples. Within the GBM group, there is evidence of heterogeneity, as some samples show slightly varied correlation patterns, reflecting the biological diversity of glioblastoma cells. These findings underscore the unique regulatory roles of miRNAs in GBM and their potential utility in distinguishing GBM-specific expression profiles from controls.

```r
# Handle row names
if (!is.null(rownames(data)) && all(!is.na(rownames(data)))) {
  # Clean row names by removing any suffix like "|0"
  rownames(data) <- str_remove(rownames(data), "\\|.*")
```

```
} else {
  data <- data %>% column_to_rownames(var = "miRNA_ID")
}

# Validate row names: Remove rows with missing or duplicate row names
data <- data %>%
  filter(!is.na(rownames(data)))  # Ensure no missing row names
rownames(data) <- make.unique(rownames(data))  # Ensure uniqueness

# Ensure all data is numeric
data <- data %>% mutate(across(everything(), as.numeric))

# Validate data: Remove rows with missing values
data <- na.omit(data)

# Transpose the data for correlation across miRNAs
data_t <- as.data.frame(t(data))
```

```
# Compute correlation matrix for miRNAs (rows)
miRNA_correlation_matrix <- cor(data_t, use = "pairwise.complete.obs")

# Convert correlation matrix to a tidy format
correlation_data <- as.data.frame(as.table(miRNA_correlation_matrix)) %>%
  filter(Var1 != Var2)  # Exclude self-correlations

strongest_correlations <- correlation_data %>%
  mutate(abs_value = abs(Freq)) %>%
  arrange(desc(abs_value)) %>%
  slice(1:25) %>%
  arrange(desc(Freq))


# Extract unique variables involved in top correlations
unique_vars <- unique(c(strongest_correlations$Var1,
                        strongest_correlations$Var2))

# Prepare the subset correlation matrix
heatmap_matrix <- miRNA_correlation_matrix[unique_vars, unique_vars]

# Create the heatmap
pheatmap(
  heatmap_matrix,
```

```
  cluster_rows = TRUE,
  cluster_cols = TRUE,
  display_numbers = FALSE,
  main = "Top 25 Strongest miRNA Correlations",
  color = colorRampPalette(c("white", "orange", "brown"))(50)
)
```



The heatmap displays the top 25 strongest correlations between miRNAs, with correlation coefficients ranging from approximately 0.9992 to 1. Hierarchical clustering organizes miRNAs into groups based on their correlation patterns, indicating potential shared regulatory mechanisms or biological functions. The dominance of warm tones across the heatmap suggests consistently high co-expression among these miRNAs, highlighting their tightly interconnected nature. This strong correlation may imply that these miRNAs are co-regulated or participate in similar pathways, making them promising candidates for further exploration in glioblastoma biology or therapeutic targeting.

```
# Summary table for correlations
correlation_summary <- correlation_data %>%
  summarise(
    Positive = sum(Freq > 0),
    Negative = sum(Freq < 0),
    Mean = mean(Freq, na.rm = TRUE),
    Median = median(Freq, na.rm = TRUE),
```

Table 2: Summary of miRNA Correlations

Positive	Negative	Mean	Median	Minimum	Maximum
261302	131200	0.4000826	0.6264028	-0.9076791	0.9998995

```
    Minimum = min(Freq, na.rm = TRUE),
    Maximum = max(Freq, na.rm = TRUE)
  )

# Display the summary table using kable
correlation_summary %>%
  kbl(caption = "Summary of miRNA Correlations") %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover",
                                                          "condensed"))
```

The summary of miRNA correlations indicates a total of 261,302 positive correlations and 131,200 negative correlations among the miRNAs. The mean correlation value is 0.40, suggesting a general tendency toward positive correlations, while the median is 0.63, highlighting a skew toward stronger positive relationships. The minimum correlation value is -0.91, showing the presence of strong negative correlations, while the maximum is nearly 1, reflecting near-perfect positive correlations. This distribution underscores the complex and diverse interactions among miRNAs, with both cooperative and antagonistic regulatory relationships evident.