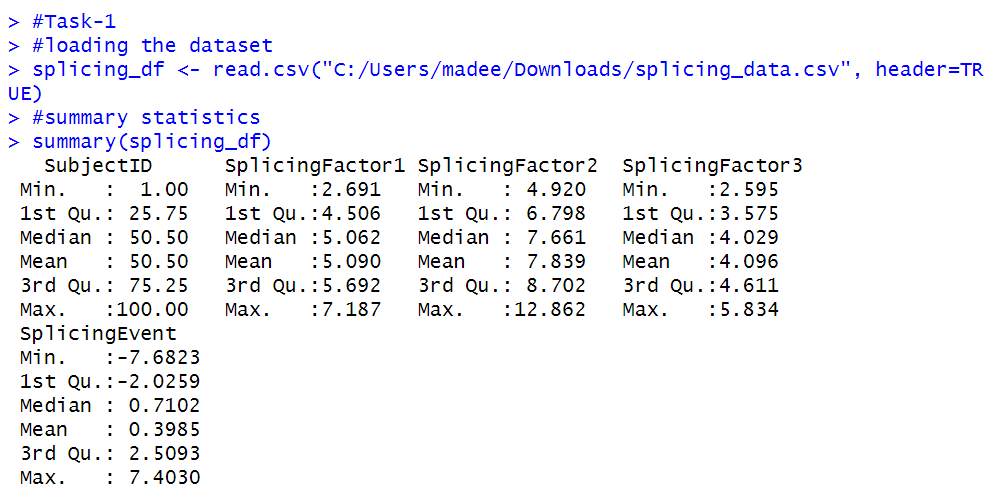
# **BINF5040E001**

**Project**

**TASK-1: DATA EXPLORATION**

**Dataset loading**

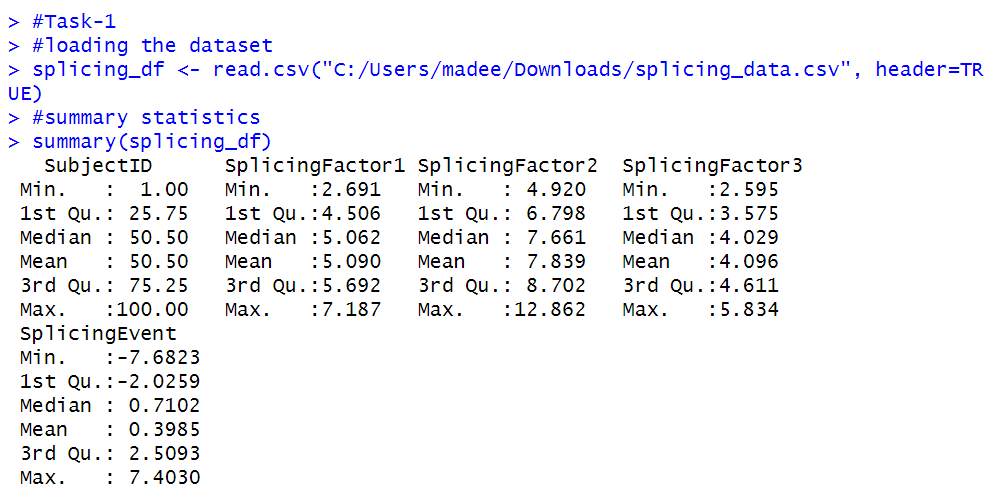


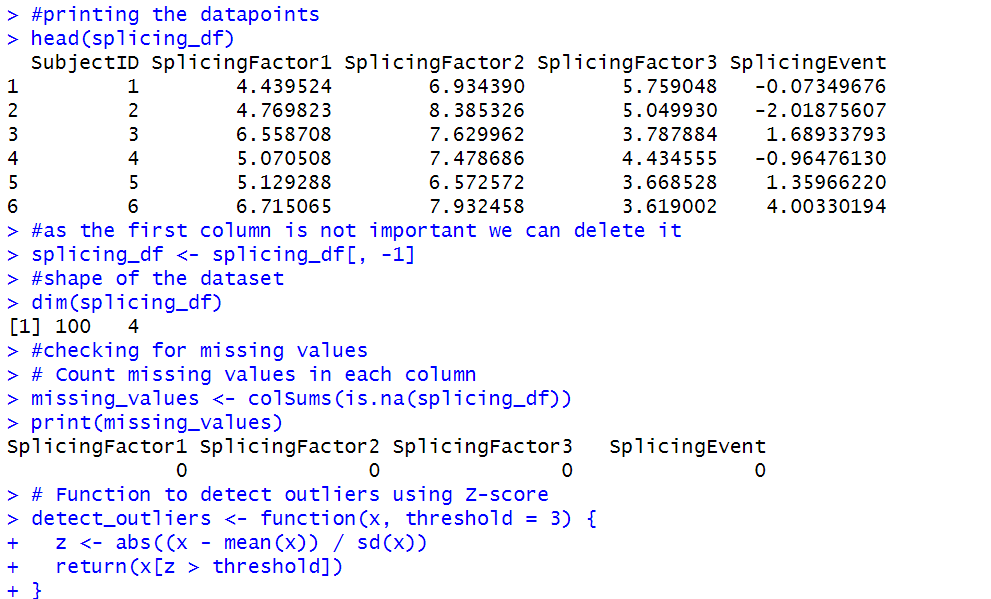


The dataset contains four features excluding the subject id.

The dimension of our dataset is (100,5).

It means we contain 100 instances of the data points and 5 features.

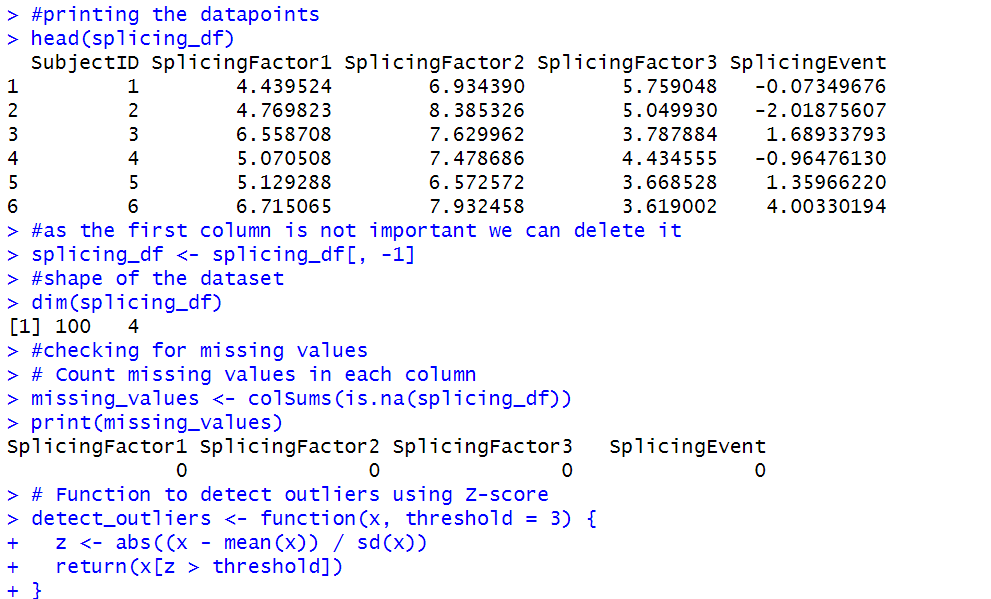
**Summary**

For the splicing factors, the summary indicates the range of values they take, as well as their distribution across the dataset. For instance, the median values for SplicingFactor2 and SplicingFactor3 are around 5.062 and 7.661, respectively, indicating that these factors tend to have values around these points in the dataset.

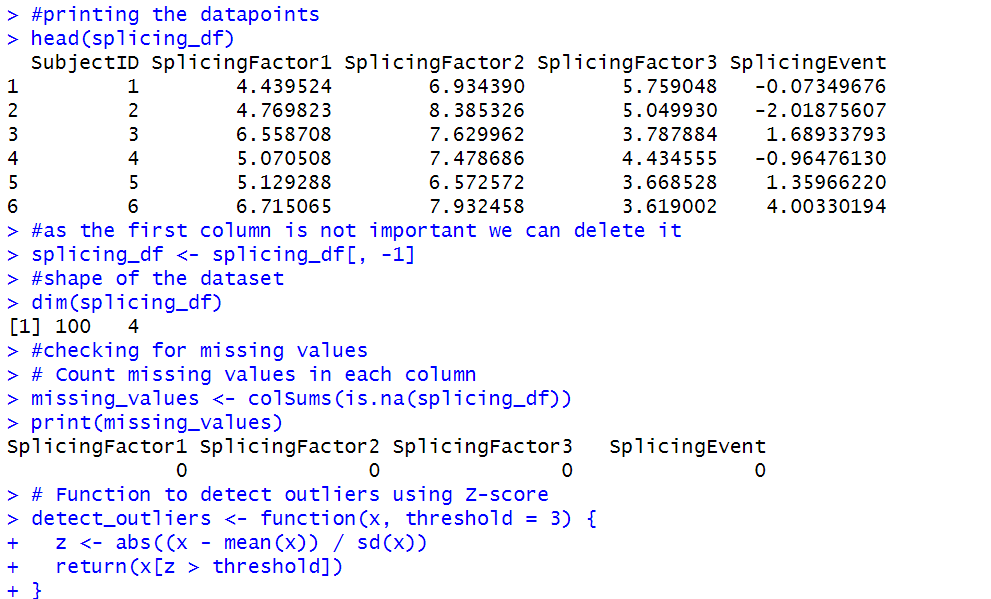
The statistics for the SplicingEvent variable suggest its distribution and central tendency. For example, the mean of 0.3985 suggests that, on average, the splicing event measure is close to zero, while the range of values (from -7.6823 to 7.4030) indicates the variability of this measure across the dataset.

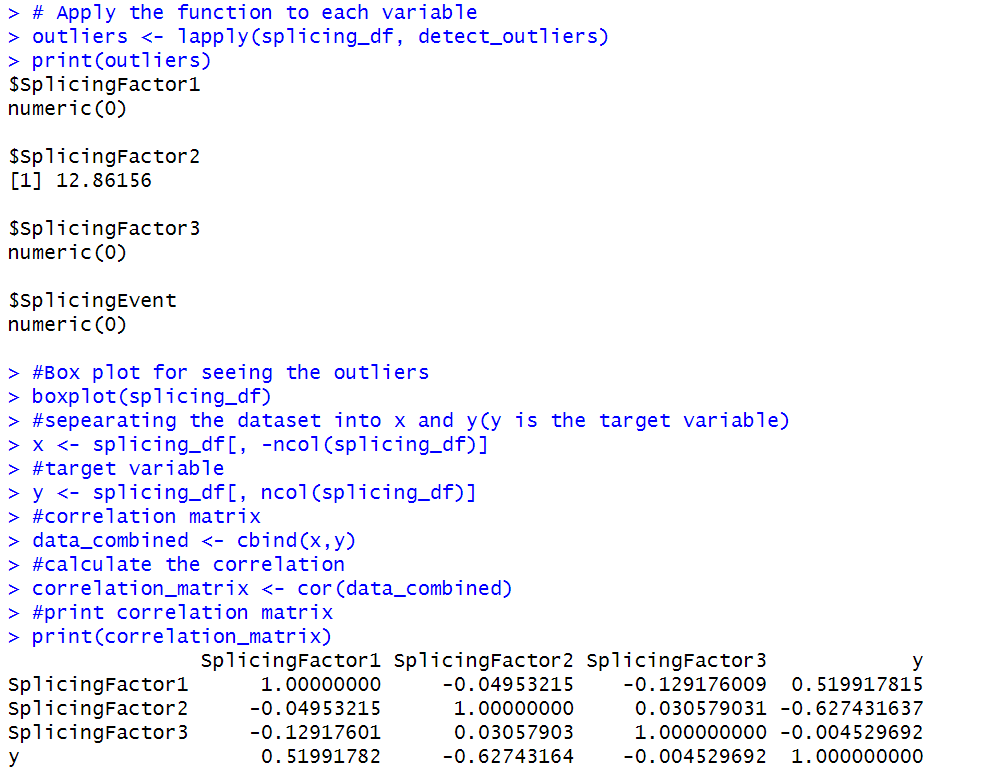
**Missing values**

our dataset is cleaned means it doesn't contain any missing values.

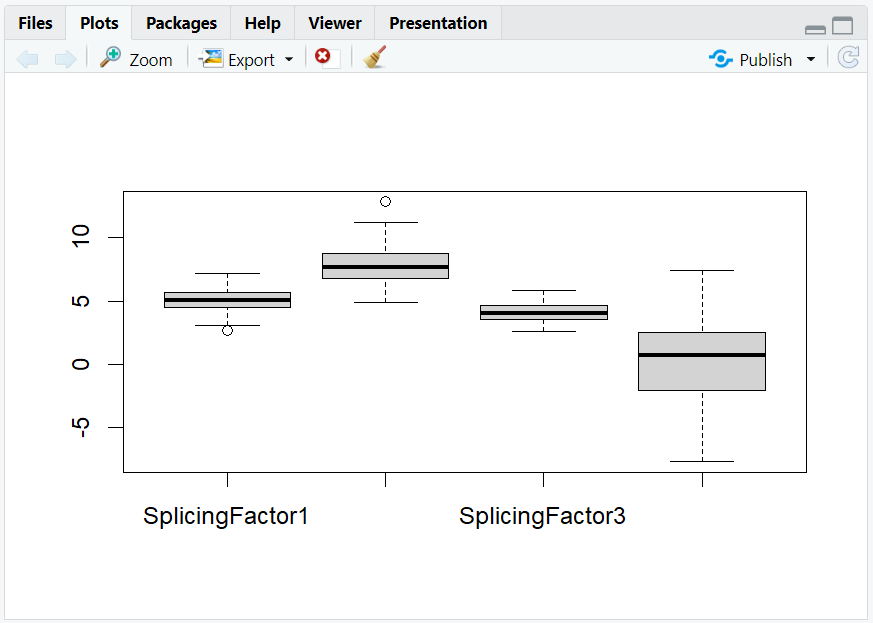


**Outliers**

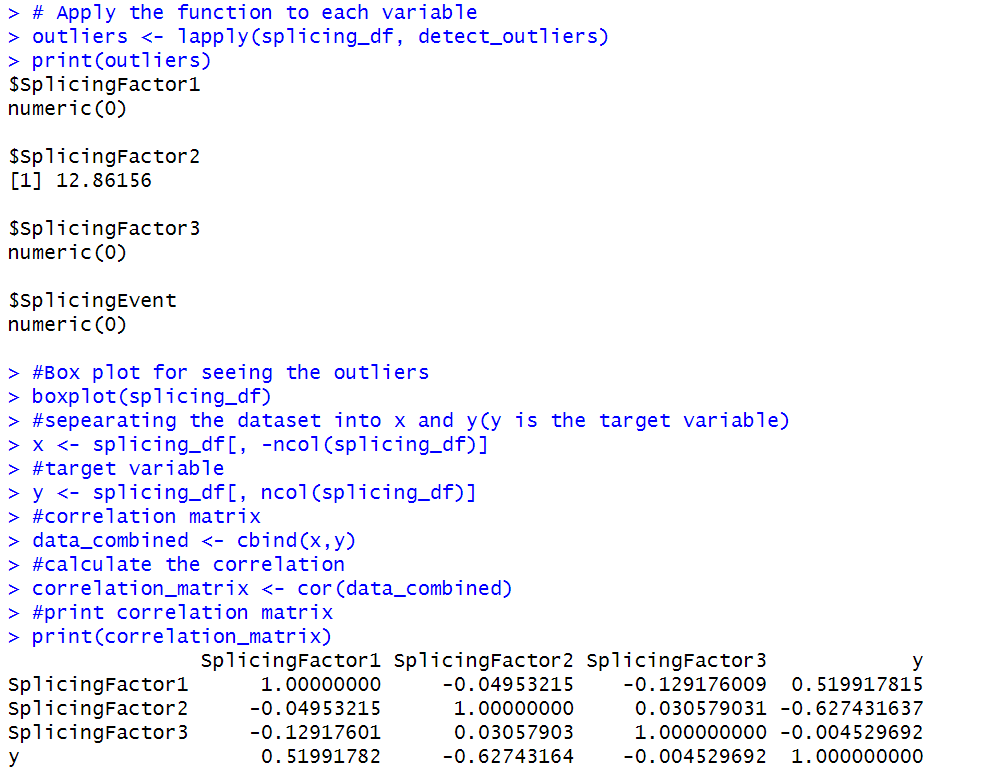




The following plot suggests that we have outlier only in the splicing factor2 column. Its value is 12.8615(there in the code file). This value is significantly different from the all the value present in that column of the dataset.



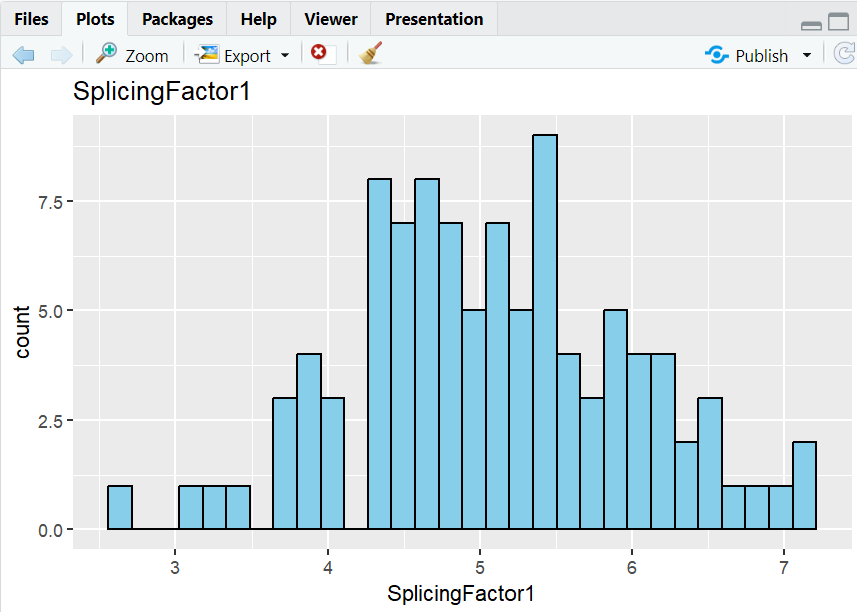
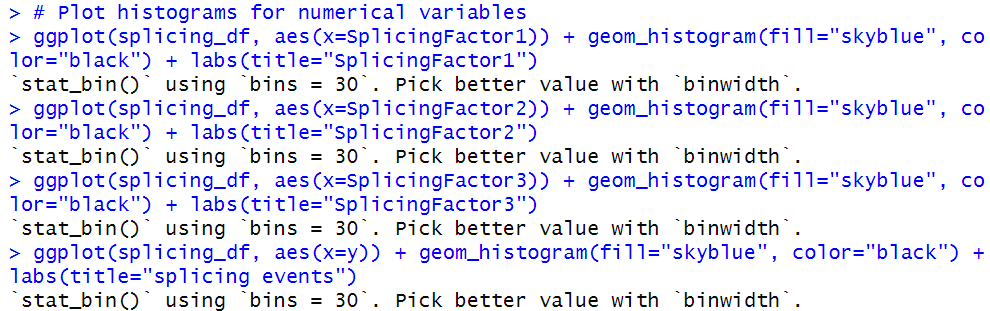
**Correlation matrix**



It has moderate positive correlation with "SplicingFactor1" (0.5199), indicating a positive relationship, and moderate negative correlation with "SplicingFactor2" (-0.6274), indicating a negative relationship. The correlation with "SplicingFactor3" is very weak (- 0.0045), suggesting little to no relationship.

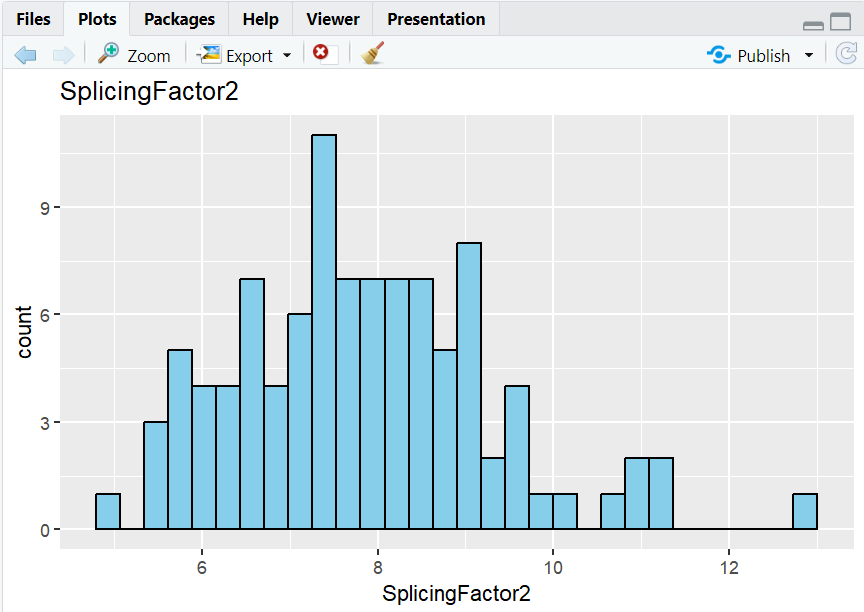
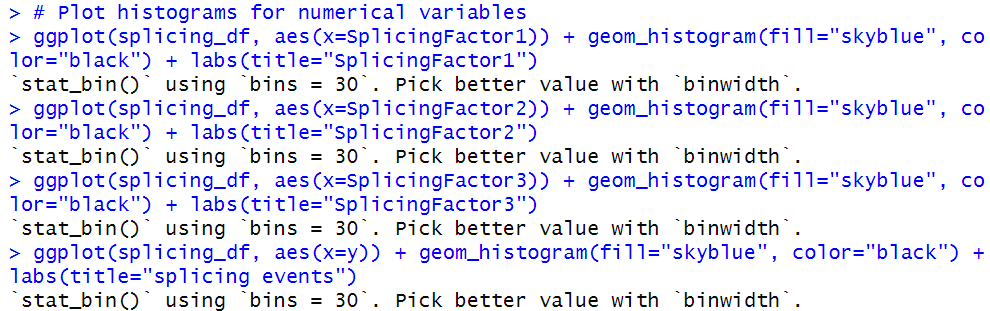
**TASK-2: DATA VISUALIZATION**

**Creating visualization to explore the distribution of each splicing factors expression and splicing events**



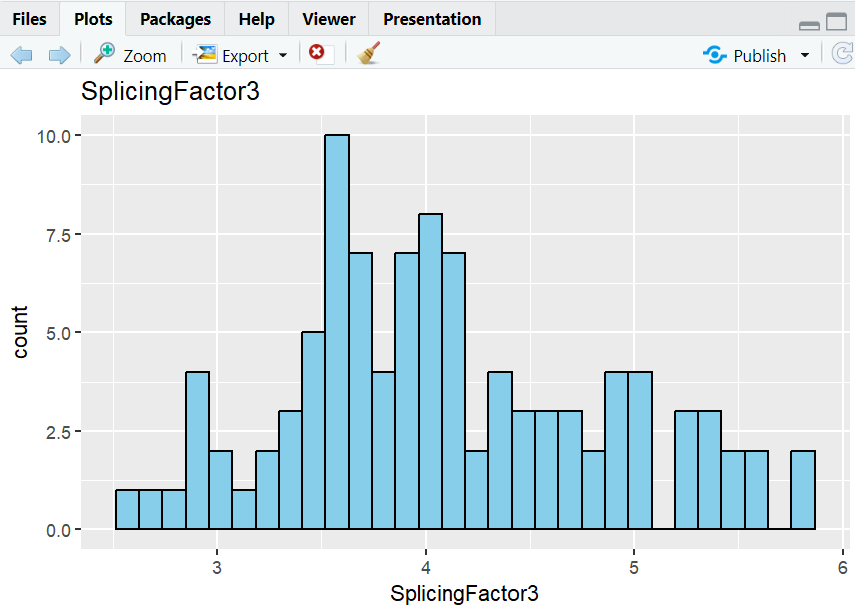
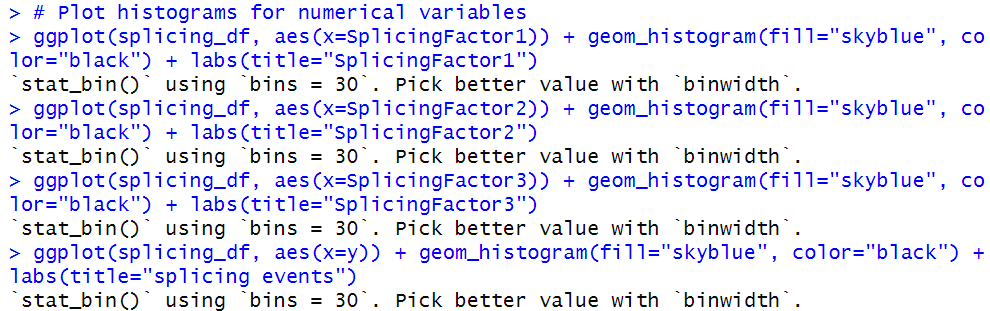
Distribution of splicingFactor1 features.

We have a greater count of splicing factor when its value increases.



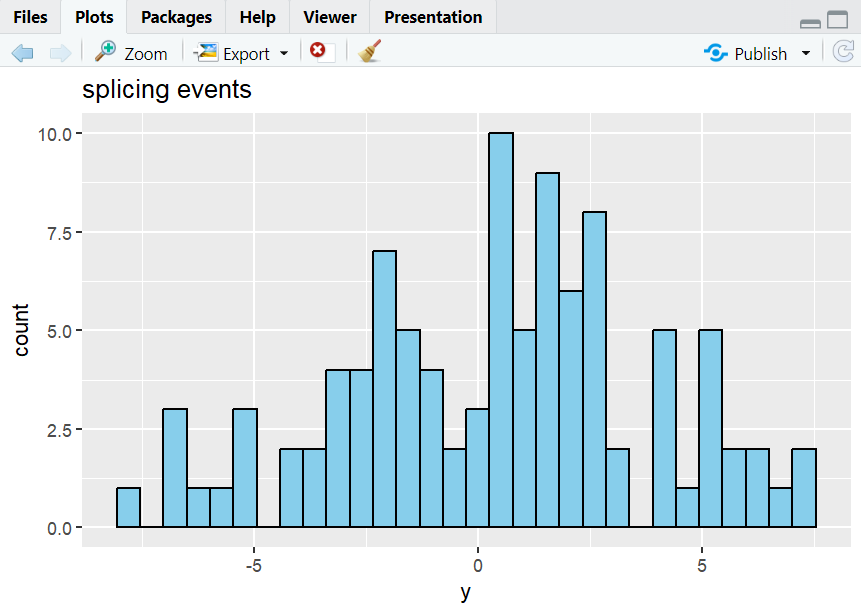
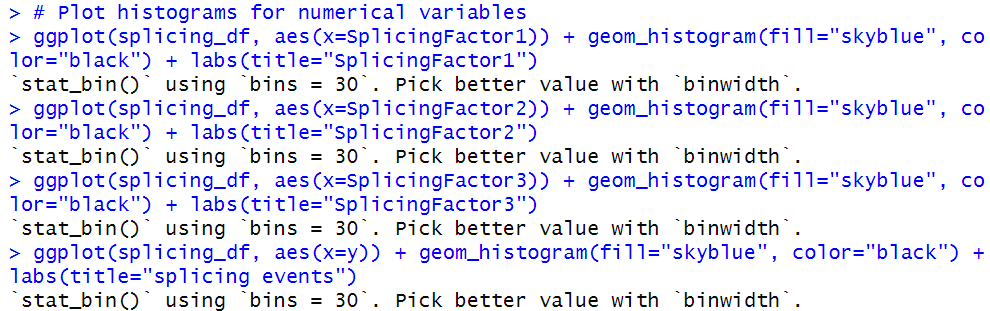
Distribution of splicingFactor2 features.

We have lesser count of splicing factor when its value increases.



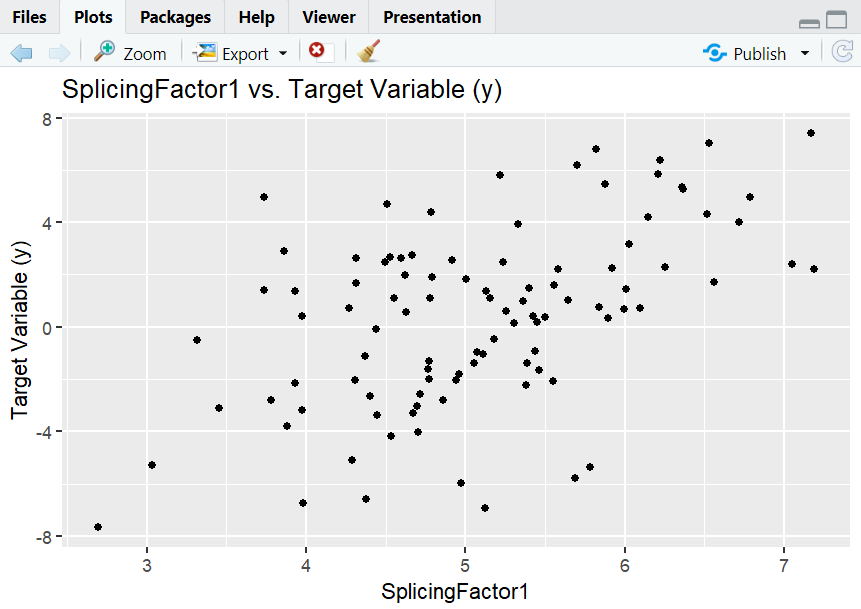
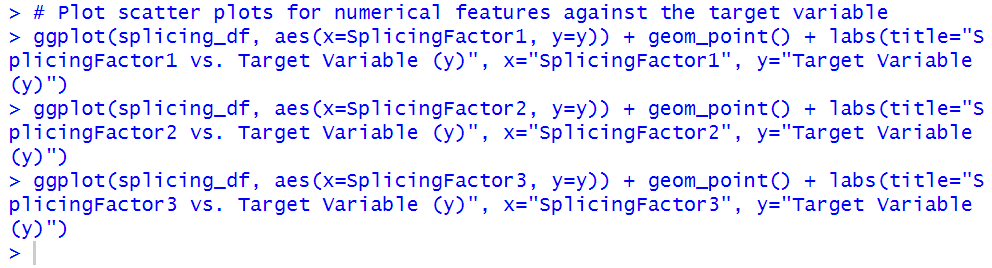
Distribution of splicingFactor3 features.

We have normalized the count of splicing factor when its value increases.



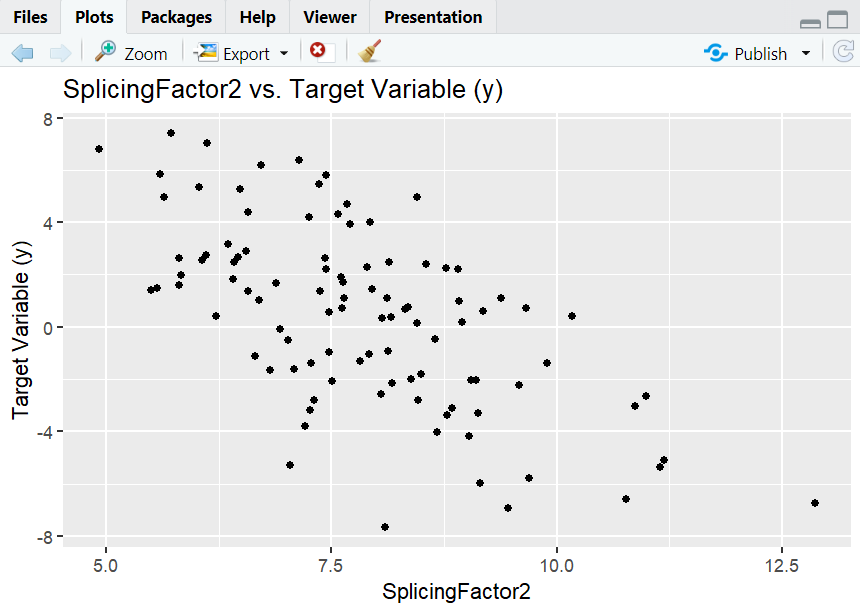
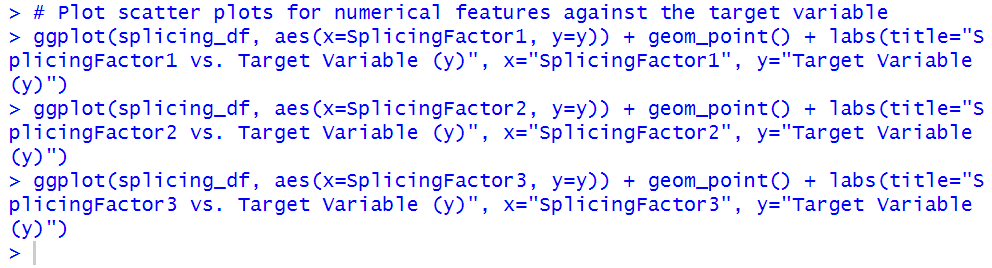
Target variable values distributions. Its greater count of value at the middle of the range of the given value of the target variable.

**Creating graphs to visualize the relationships between each individual splicing factor and the splicing event**



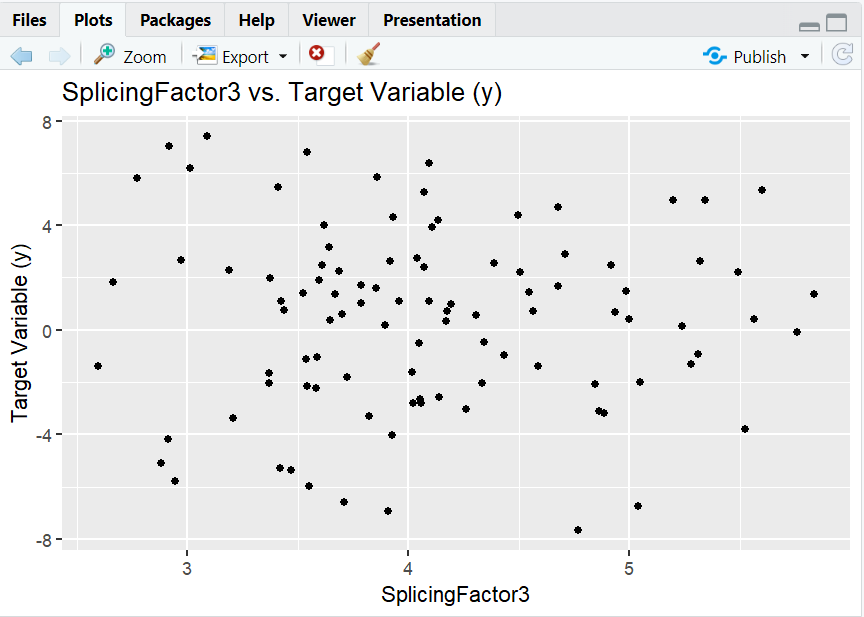
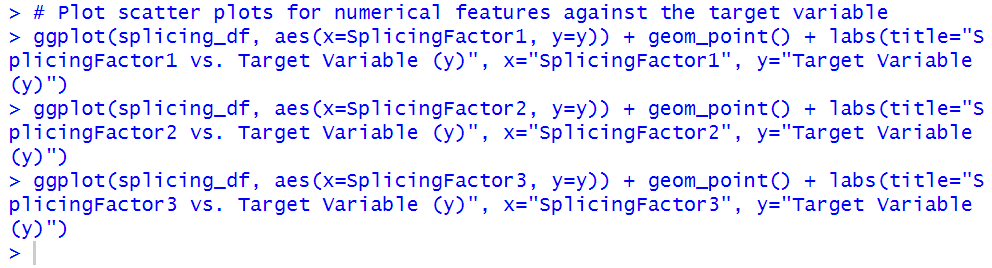
In between the splicing factor1 and target variable.

It suggests that the relationship is linear in nature. As the value of splicing factor increases the target variable value also increases.



In between the splicing factor2 and target variable.

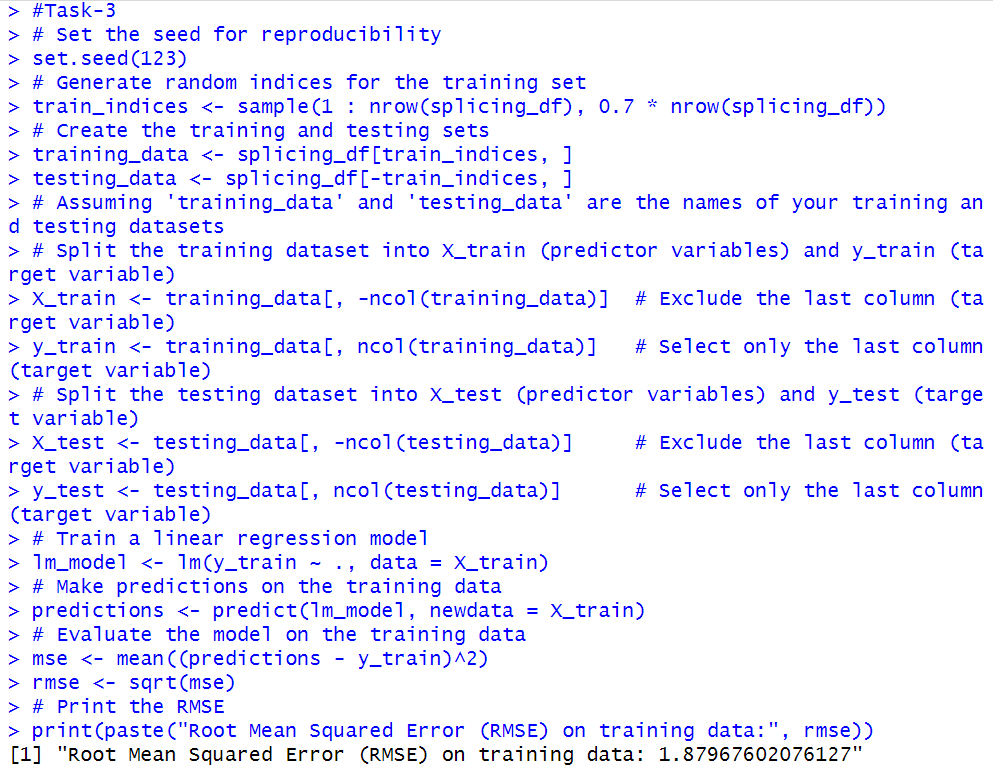
It suggests that the relationship is linear in nature. As the value of splicing factor increases the target variable value decreases.

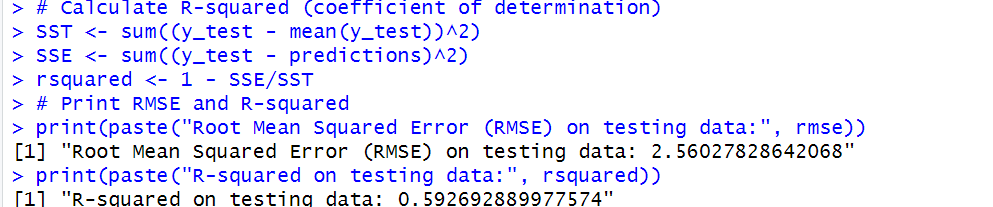
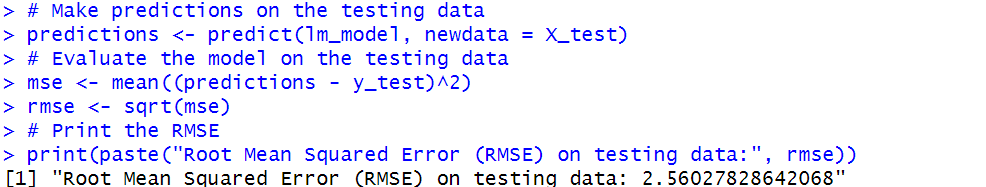


In between the splicing factor3 and target variable.

It suggests that the relationship is not linear in nature. The data points are scattered across the graph.

**TASK-3: PREDICTIVE MODELING**





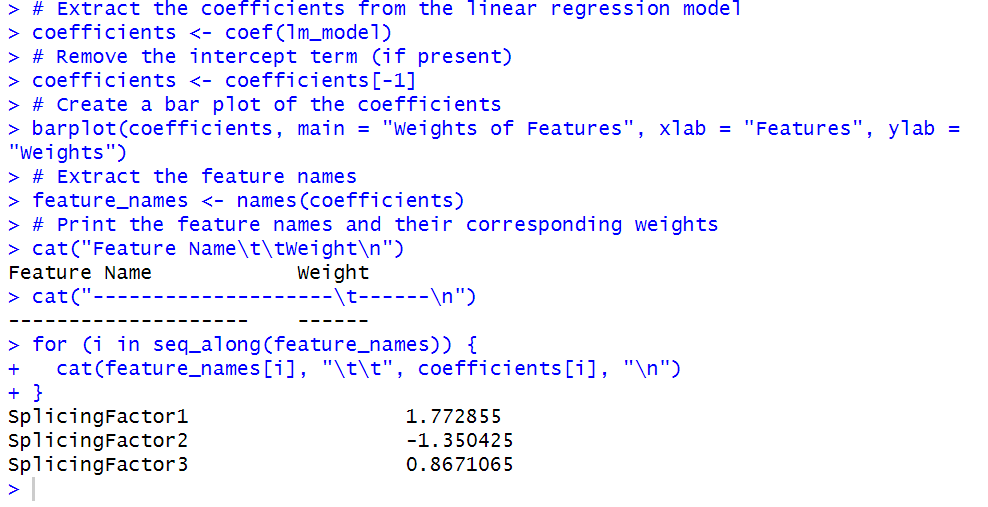
**Root Mean Squared Error (RMSE):**

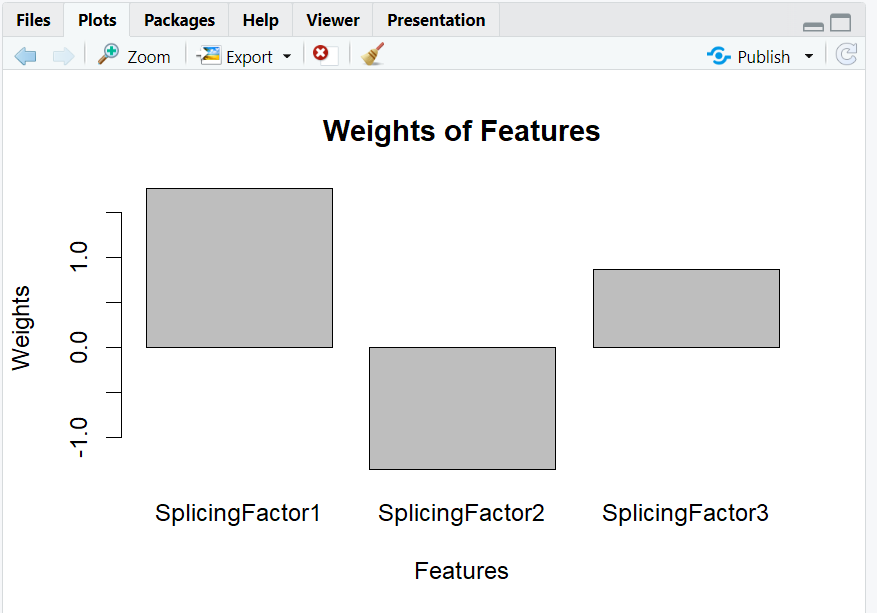
RMSE value is 2.560, which represents the square root of the average squared difference b/w the predicted values and the actual values (target variable) in the testing data. In other words, on average, the predictions made by the linear regression model on the testing data are approximately 2.560 units away from the actual values.

**R-squared (Coefficient of Determination):**

R-squared value is 0.593, which indicates the proportion of the variance in the target variable (y\_test) that is explained by the independent variables (X\_test) in the model. In this case, approximately 59.3% of the variance in the target variable is explained by the linear regression model, while the remaining 40.7% of the variance is unexplained or attributed to random error. Higher R-squared values closer to 1 indicate that the model fits the data well, while lower values suggest that the model may not capture all the variability in the data. In summary, these values provide insights into the performance of the linear regression model on the testing data. The RMSE quantifies the average error of the model's predictions, while R-squared indicates how well the model explains the variability in the target variable.

**Interpretation of coefficients:**





The weights of each parameter are clearly represented by the bar plot, which facilitates simple comparison of their relative importance.

**Interpretation:**

The influence of features is represented by weights: The weights (coefficients) in the linear model show the relative contribution of each feature to the expected result.

Features with positive weights have a positive correlation with the result, whereas those with negative weights have a negative correlation. Positive weights rise, negative weights fall.

Greater absolute weight equals stronger impact: A feature's effect increases with its absolute value.

**Specific insights**

* With the highest positive weight, SplicingFactor1 is predicted to have the most positive correlation with the expected result.
* SplicingFactor2's negative weight suggests a poor correlation with the result.
* SplicingFactor3 is weaker than SplicingFactor1 but has a favorable weight: It has a less noticeable effect than SplicingFactor1.

**TASK-4: CONCLUSIONS & RECOMMENDATIONS**

All the key findings are being displayed above. In the respective sections.

**Prior feature:**

The cofficients(weights) plot of the above gives us certain idea.

* The splicing factor1 has a weight of approximately 1.773. since its positive in nature and indicates a positive relationship with the target variable.
* The splicing factor 2 had a negative relationship with the target variable.
* The last feature had a positive relationship with the target variable.

Comparing the absolute values of the weights, we can see that SplicingFactor1 has the strongest relationship with the target variable among the three features, as it has the largest absolute weight.

**Recommendation:**

Any model requires a large amount of data points to learn general patterns or trends in those data points. Always try to get more and more data points.

We can use some other algo to get the result.