**House Pricing Prediction**   
Team 70

Youssef Khaled Abdallah 20191700767

Youssef Omar Hussein 20191700776

Youssef Khaled Farouk 20191700768

# Required Libraries

1. **Caret**.
2. **RandomForest**.
3. **GBM**.

# Preprocessing

1. **ExploreData(data)**

displays the number of columns, number of rows, count of missing values, and summary statistics of the data

1. **RemoveNullColumns(threshold, train\_data, test\_data)**

Drop columns that have more than a **threshold** percentage of **NULL** values. It calculates the **NULL** percentage for each column, identifies columns exceeding the threshold, and removes those columns from the train and test datasets.

1. **EncodeCategoricalData(train\_data, test\_data)**

Performs **one-hot encoding** on **categorical variables** in the train and test datasets. It saves the target variable from the train dataset, removes the target variable, combines train and test datasets, applies **one-hot encoding**, and **normalizes** the data using **Z-scores**. It splits the encoded data back into train and test datasets.

1. **FillNulls(data)**

Replace **NULL** values in a dataset with the median value of each respective column.

1. **Preprocess(train\_data, test\_data)**

Integrate the preprocessing steps by calling the **RemoveNullColumns**, **EncodeCategoricalData**, and **FillNulls** functions. It returns the preprocessed **x\_train, x\_test, y\_train, y\_test, train\_ids, and test\_ids.**

1. **TrainValidateSplit(x, y, ids, splitSize=0.8, seed=100)**

Shuffles and splits the preprocessed train dataset into training and validation sets. It adds IDs to the datasets, randomly partitions the data based on a specified split size and returns the split datasets.

1. **RemoveLowCorrelation(x\_train, y\_train, threshold)**

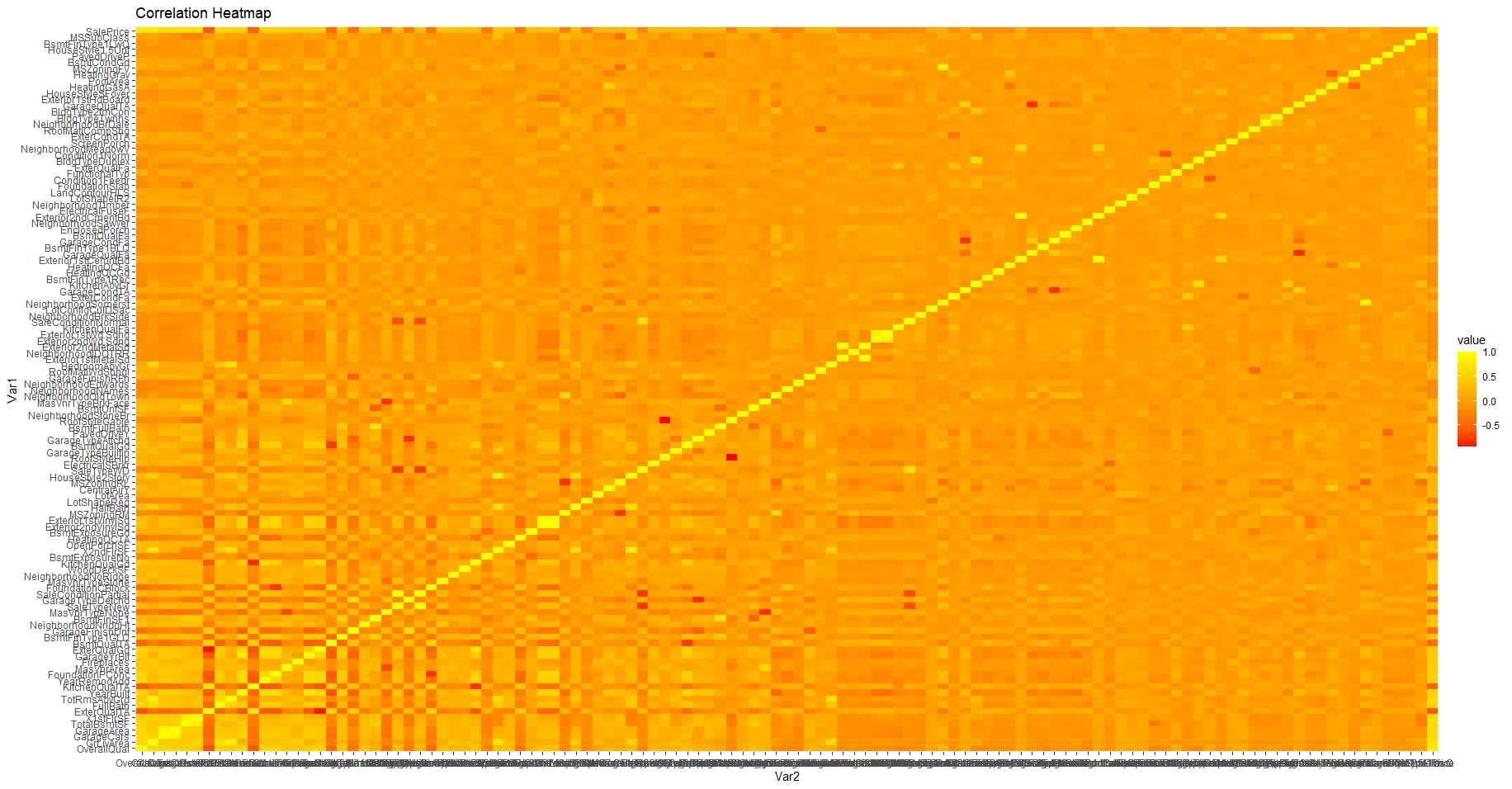
Removes Least **threshold** percentage correlation columns

Model

**Gradient Boosting machine :** It is a type of ensemble method that combines multiple weak prediction models (typically decision trees) to create a strong predictive model.

# Data Visualization

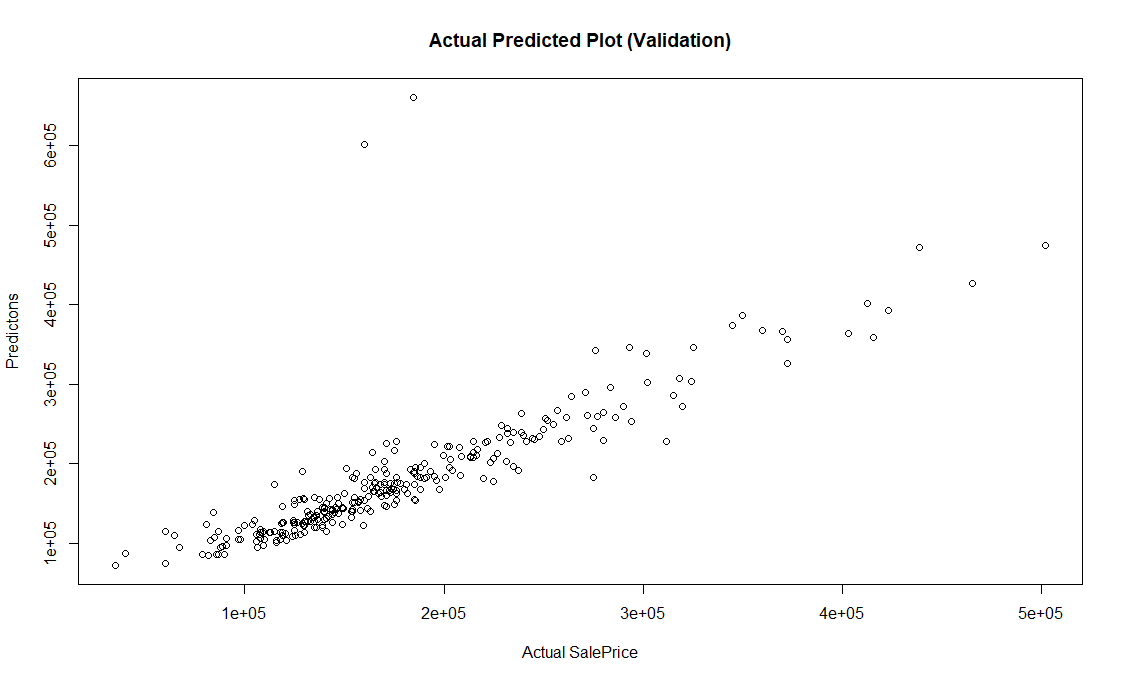
**Variables Correlations Heatmap**



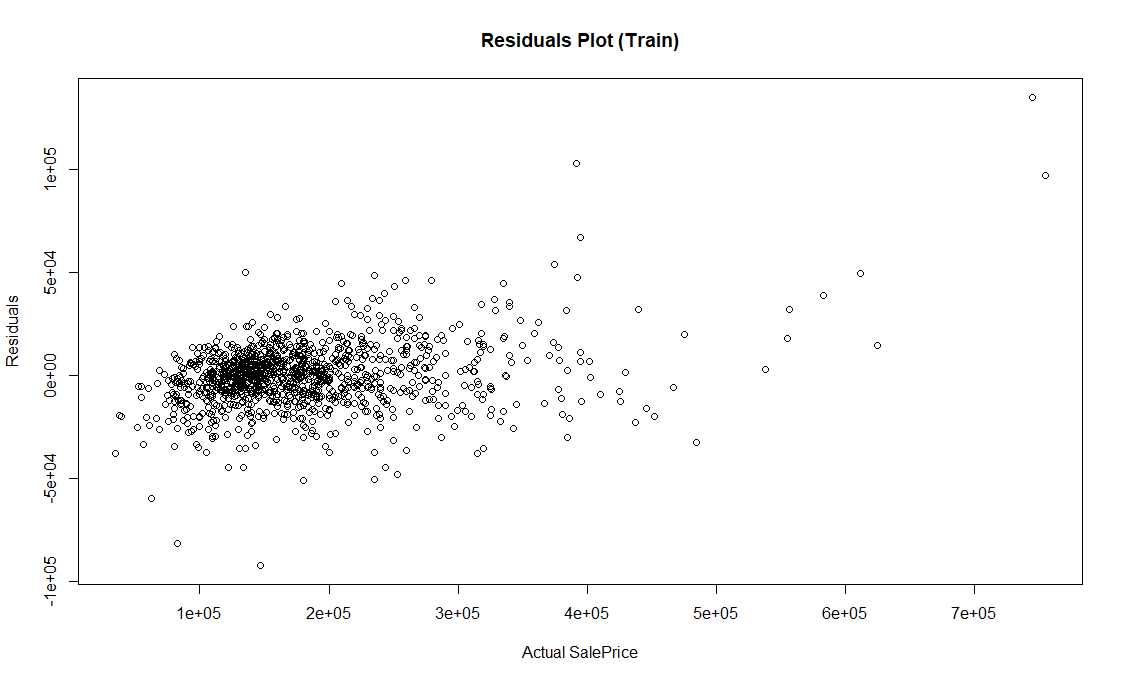
**Actual vs Predicted Plot:**

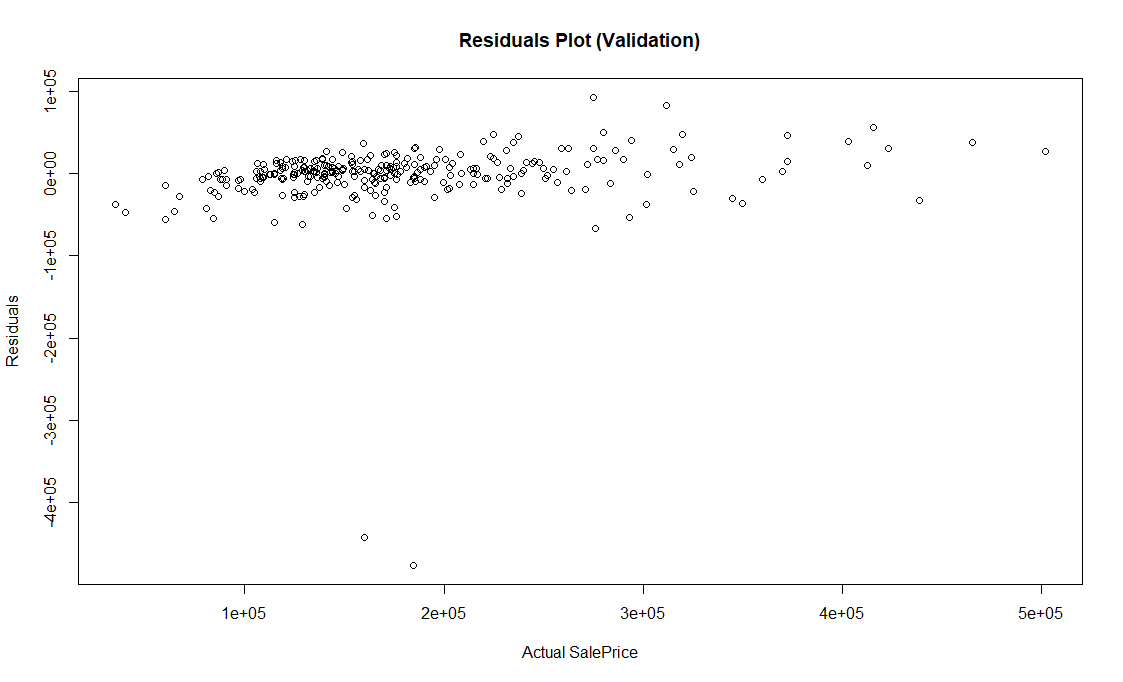
**A picture containing text, diagram, line, screenshot

Description automatically generated**

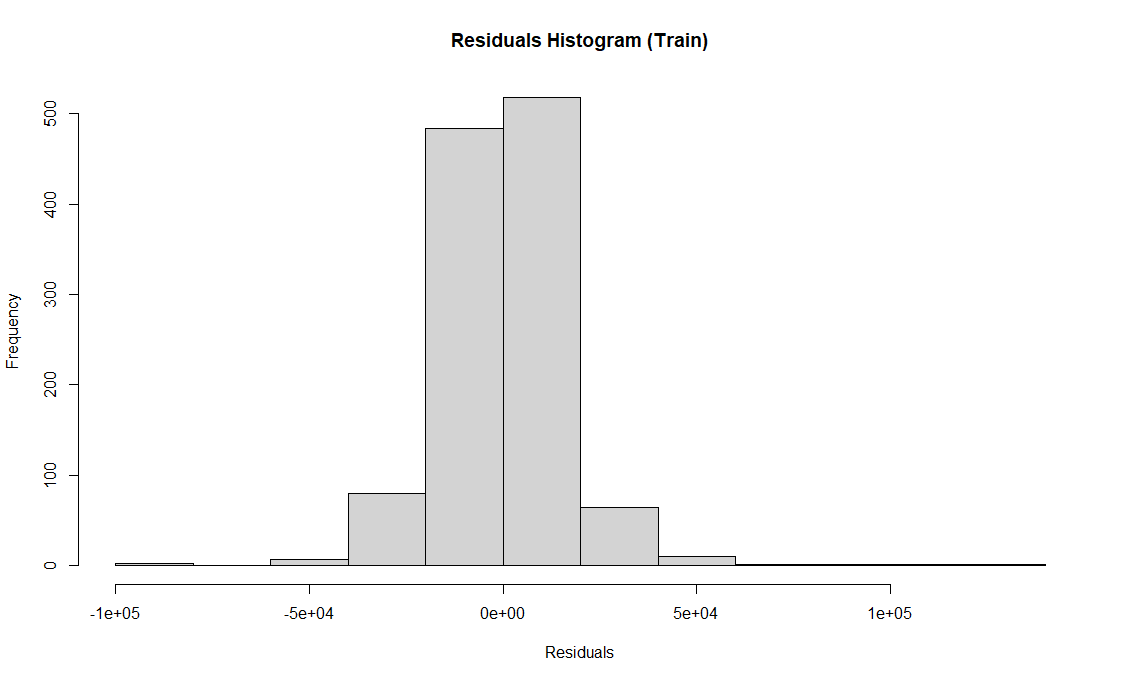
****

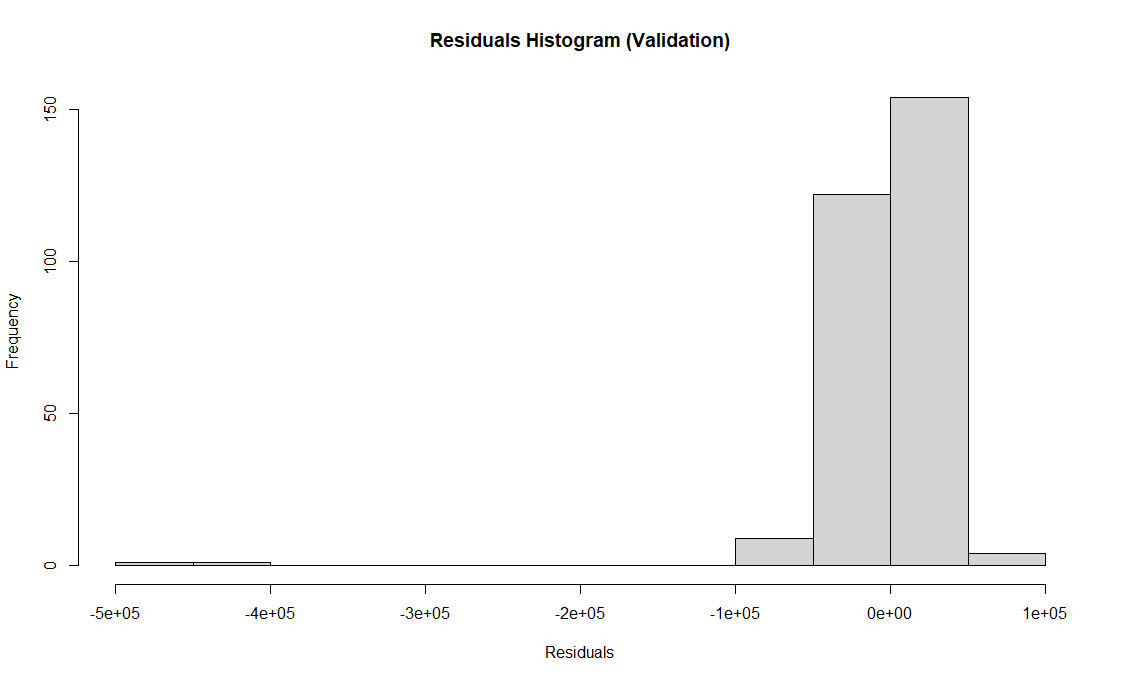
**Residuals Plot:**

****

****

**Residuals Histogram:**

****



# Evaluation Metrics

The evaluation metrics for the **training dataset** are as follows:

* Mean Squared Error (MSE): 249,788,746
* Root Mean Squared Error (RMSE): 15,804.71
* R-squared (R²): 0.9619353

The evaluation metrics for the **validation dataset** are as follows:

* Mean Squared Error (MSE): 1,912,787,428
* Root Mean Squared Error (RMSE): 43,735.43
* R-squared (R²): 0.710436

# Summary

We have successfully applied preprocessing techniques, data visualization, utilized the Gradient Boosting Machine model, and evaluated the performance of house pricing prediction model using various metrics.