STEP 5  
The analysis of the model’s performance based on the provided visualizations and statistical metrics allows us to delve into a comprehensive evaluation as part of our study:

### Residual Distribution Analysis

#### Histogram of Residuals

The distribution of residuals, as illustrated in the histogram, closely resembles a normal distribution with a pronounced peak around zero. This indicates that the model's errors are symmetrically distributed, suggesting a good fit for the majority of predictions. However, the tails seem slightly heavier than those of a perfect normal distribution, indicating the presence of outliers or extreme values that the model did not predict accurately.

#### Residual Plot Analysis

The residual plot shows residuals plotted against the actual values, where ideally all points should randomly scatter around the horizontal line at zero (no error). The plot does not exhibit any clear pattern such as a funnel shape or systematic increase/decrease along the actual value axis, which is favorable as it suggests no obvious non-linearity or heteroscedasticity issues. However, the cloud of points does exhibit some spread, hinting at variability in the prediction errors across the range of actual values. This could be an indication of model variance or certain data points that are particularly challenging to predict.

### Statistical Metrics Analysis

- \*\*Mean Squared Error (MSE)\*\*: The MSE of 15842.48 is relatively high, which can be attributed to the influence of squared large errors due to possible outliers or extreme values in the dataset.

- \*\*Root Mean Squared Error (RMSE)\*\*: An RMSE of 125.87, being the square root of MSE, provides a more interpretable measure of average error in the same units as the target variable. Given the scale of the target variable values (as seen from the range in the residual plot), this RMSE can be considered relatively low, suggesting good model accuracy.

- \*\*Mean Absolute Error (MAE)\*\*: The MAE of 88.45 is less sensitive to extreme values than MSE, providing a measure of the average magnitude of the errors. The relatively lower value of MAE compared to RMSE further highlights the presence of some large errors skewing the MSE.

- \*\*Coefficient of Determination (R²)\*\*: An R² of 0.9857 indicates that approximately 98.57% of the variance in the dependent variable is predictable from the independent variables. This suggests an excellent fit of the model to the data.

### Identified Issues and Recommendations

The analysis indicates a potential issue of heteroscedasticity, as suggested by the varying spread in residuals across the range of actual values, though it isn’t prominently visible in the residual plot provided. Heteroscedasticity involves non-constant variance of error terms, violating an assumption of linear regression models that could lead to inefficient estimates.

To address this:

- \*\*Transforming the Target Variable\*\*: Applying transformations such as logarithmic or square root on the target variable can help stabilize variance across the data.

- \*\*Using Weighted Regression Techniques\*\*: Applying weights to the regression can give more importance to certain regions of data to combat heteroscedasticity.

### Conclusion

In conclusion, our model achieves a high coefficient of determination (R²), suggesting strong predictive performance. The project’s success criterion based on R² was clearly met. However, we observed indications of potential model limitations such as heteroscedasticity, recommending further investigation and possible adjustments to the model or data to enhance prediction reliability and model robustness. This detailed examination provides vital insights into the model's capabilities and areas for improvement, aligning with the overall project goals for high predictive accuracy and model interpretability.  
  
STEP 4:  
In our study, we leveraged the RandomizedSearchCV approach to conduct hyperparameter tuning on the XGBoost algorithm, favoring it over grid search due to its efficiency in exploring a vast parameter space without the exhaustive computation. The hyperparameters tuned included max depth, n\_estimators, learning\_rate, subsample, colsample\_bytree, min\_child\_weight, and gamma, which are critical in controlling the complexity and fitting behavior of the model. This tuning was executed through 50 iterations with 3-fold cross-validation, optimizing for the best parameter combination that minimized the mean squared error (MSE). Upon identifying the optimal model settings, we assessed its performance on a scaled test set, calculating metrics such as MSE, root mean squared error (RMSE), and the coefficient of determination (R²) to evaluate prediction accuracy and the variability explained by the model. A visual representation was also created to compare the actual versus predicted values, using distinctive markers and line styles to enhance clarity and interpretability of the predictive performance. This phase underscored the XGBoost algorithm's robustness and adaptability in managing complex modeling tasks, albeit highlighting the inevitable trade-offs between computational demand and accuracy, a typical challenge in advanced machine learning applications.