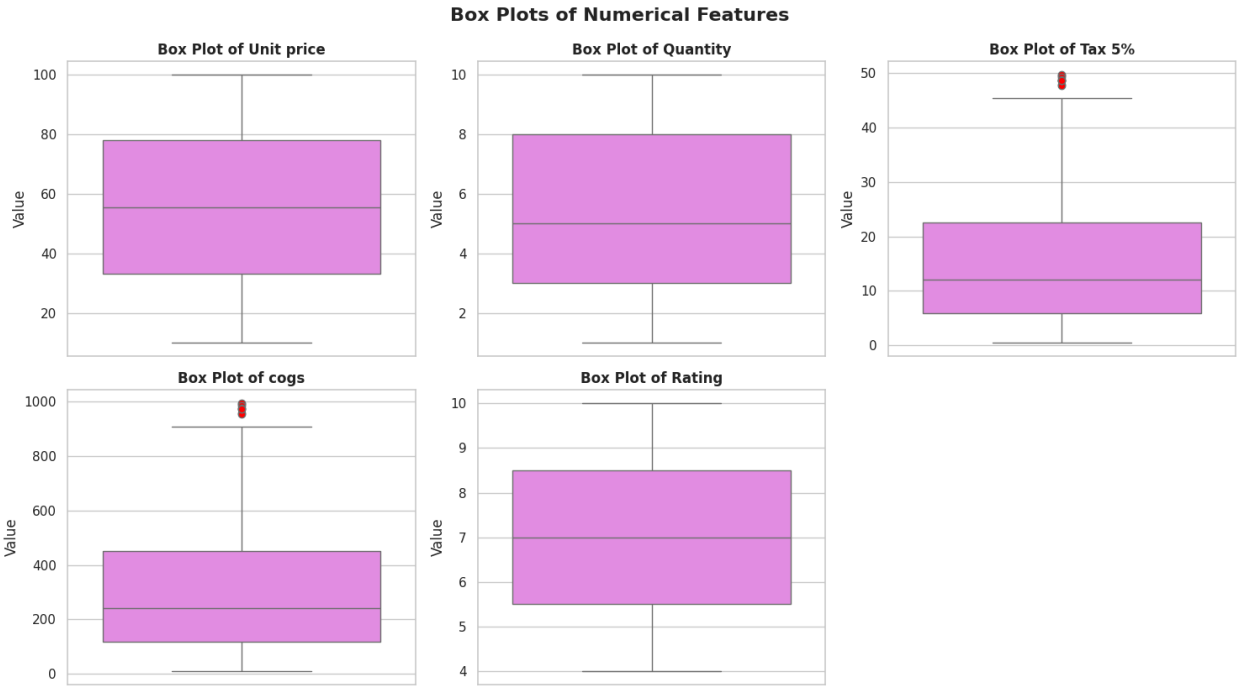

Fig-1: Box Plot Analysis of Numerical Features

Explanation:

Box Plot Analysis of Numerical Features

1. Unit Price

For the Unit price feature, the first quartile (Q1) is 33.1, and the third quartile (Q3) is 77.935, resulting in an Interquartile Range (IQR) of 44.835. The calculated lower bound for detecting outliers is -34.15, and the upper bound is 145.19. Since all the values in this column fall within these bounds, no outliers are detected. This indicates a relatively well-distributed range without extreme deviations.

2. Quantity

In the Quantity feature, Q1 is 3.0, and Q3 is 8.0, leading to an IQR of 5.0. The lower bound is -4.5, and the upper bound is 15.5. Here as well, all values lie within the permissible range, and no outliers are detected. The data is tightly clustered between the lower and upper bounds, reflecting a consistent distribution of quantity values.

3. Tax 5%

The Tax 5% feature has Q1 at 5.92 and Q3 at 22.51, giving an IQR of 16.59. The lower bound for outliers is -18.96, and the upper bound is 47.40. However, several values exceed the upper bound, which makes them outliers. These outliers are observed at indices such as 166, 167, 350, 357, and 422, with values ranging from 47.72 to 49.65. This suggests that certain transactions had significantly higher tax values compared to the majority.

4. COGS (Cost of Goods Sold)

For COGS, Q1 is 119.71, and Q3 is 450.31, leading to an IQR of 330.595. The lower bound is calculated as -376.18, and the upper bound is 946.20. Outliers are detected beyond the upper bound, with values like 955.8, 989.8, 993.0, and similar, appearing at indices like 166, 167, 350, 357, and 557. These high COGS values reflect transactions involving significantly larger costs, potentially due to bulk purchases or high-value items.

5. Rating

The Rating feature shows Q1 as 5.5 and Q3 as 8.5, resulting in an IQR of 3.0. The lower bound is 1.0, and the upper bound is 13.0. All values fall within this range, and no outliers are detected. This suggests that the ratings are generally well-behaved, with most values between 5.5 and 8.5, reflecting a positive and controlled spread.

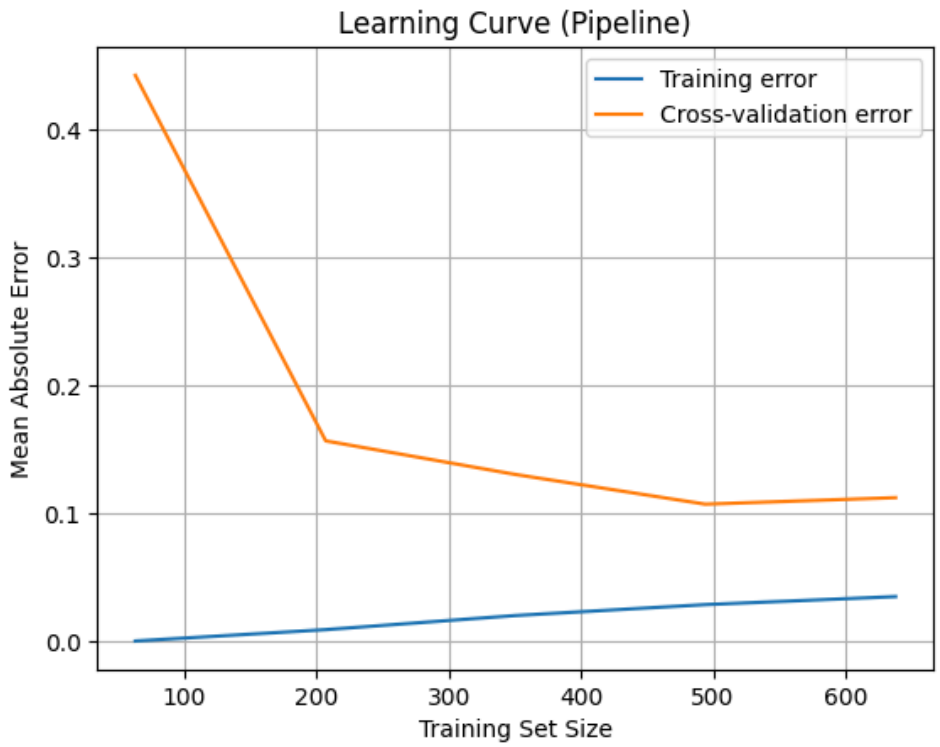

Fig-2: Before Improving GradientBoosting Learning Curve

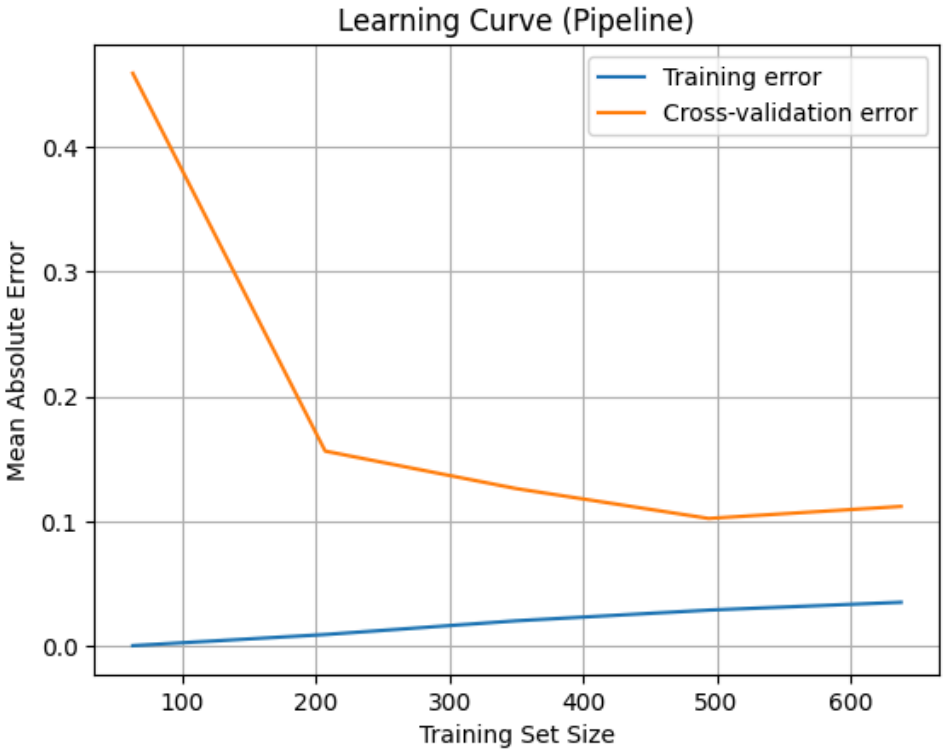

Fig-3: After Improving GradientBoosting Learning Curve

Explaination:

The two learning curves show how the GradientBoosting model performs before and after improvements as the amount of training data increases. Fig-1, the first plot (before improvement), the training error starts very low, which means the model fits the training data well. However, the cross-validation error starts very high and decreases as more data is added, but it levels off around 0.1. The large gap between training and cross-validation errors shows the model is overfitting—it memorizes the training data but struggles to perform well on new data.

Fig-2, the second plot (after improvement), the training error still increases slowly as more data is added, but the cross-validation error drops further and stays lower. Most importantly, the gap between training and cross-validation errors is much smaller now. This shows that the model is generalizing better—it's not just memorizing the training data but learning patterns that work for new data too.

Overall, the second plot shows a clear improvement. The smaller error gap and lower cross-validation error mean the model is more balanced and accurate, likely due to changes like tuning hyperparameters, adding regularization, or selecting better features. The model now performs more reliably on unseen data.

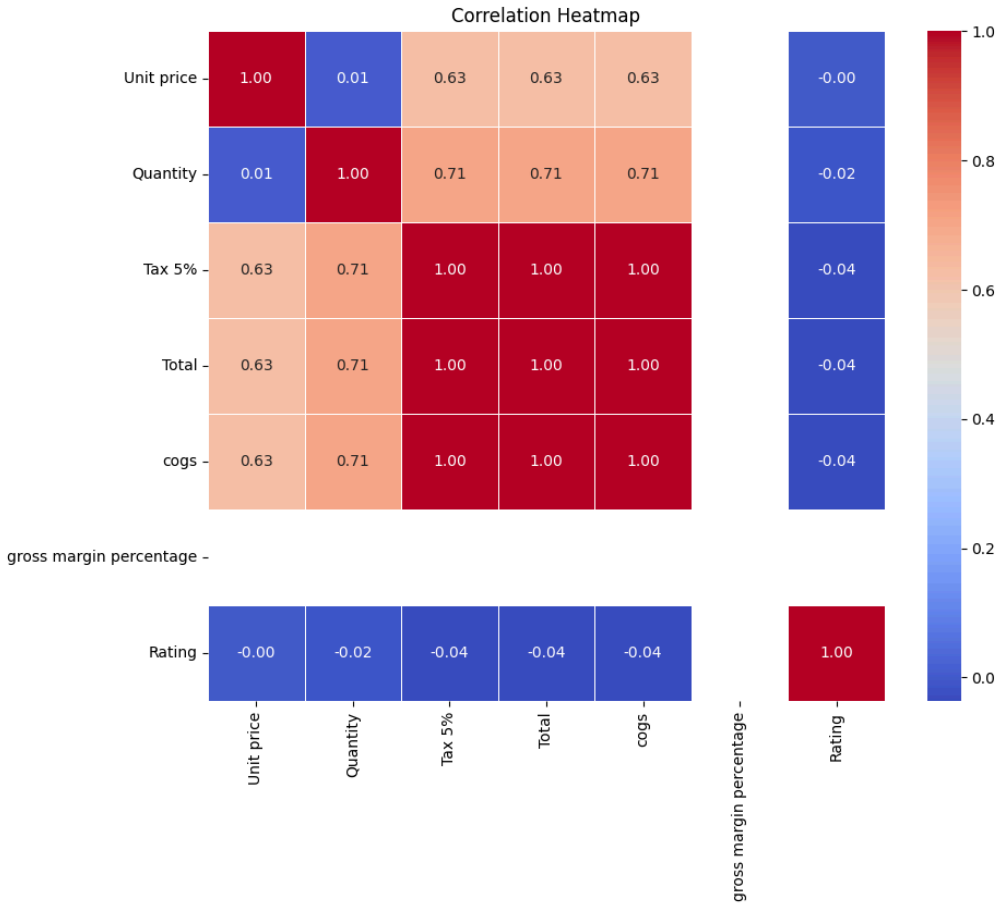

Fig-4: Numerical Feature Correlation Heatmap

Explaination:

This correlation heatmap visually shows the relationships between different numerical features in a dataset. The values range between -1 and 1, where 1 indicates a perfect positive correlation, 0 means no correlation, and negative values indicate a negative relationship.

For example, the Tax 5%, Total, and cogs are perfectly correlated with each other (all having values of 1), meaning they move exactly in sync. This makes sense because Total and cogs (cost of goods sold) naturally include Tax, so they are closely related. On the other hand, the Quantity has a moderately positive correlation (0.71) with Tax, Total, and cogs. This suggests that as the quantity increases, these values also tend to increase.

Unit Price has a weaker positive correlation (0.63) with Tax, Total, and cogs, showing that unit price also impacts these values, but not as strongly as quantity. Interestingly, the Rating has almost no correlation (close to 0) with any other features, meaning customer ratings do not depend on price, tax, or quantity.

Lastly, the feature gross margin percentage seems disconnected here, because it is not included in the correlation matrix or was likely dropped due to missing or non-numerical values in the dataset. A correlation heatmap only includes numerical features that have valid values for all observations.

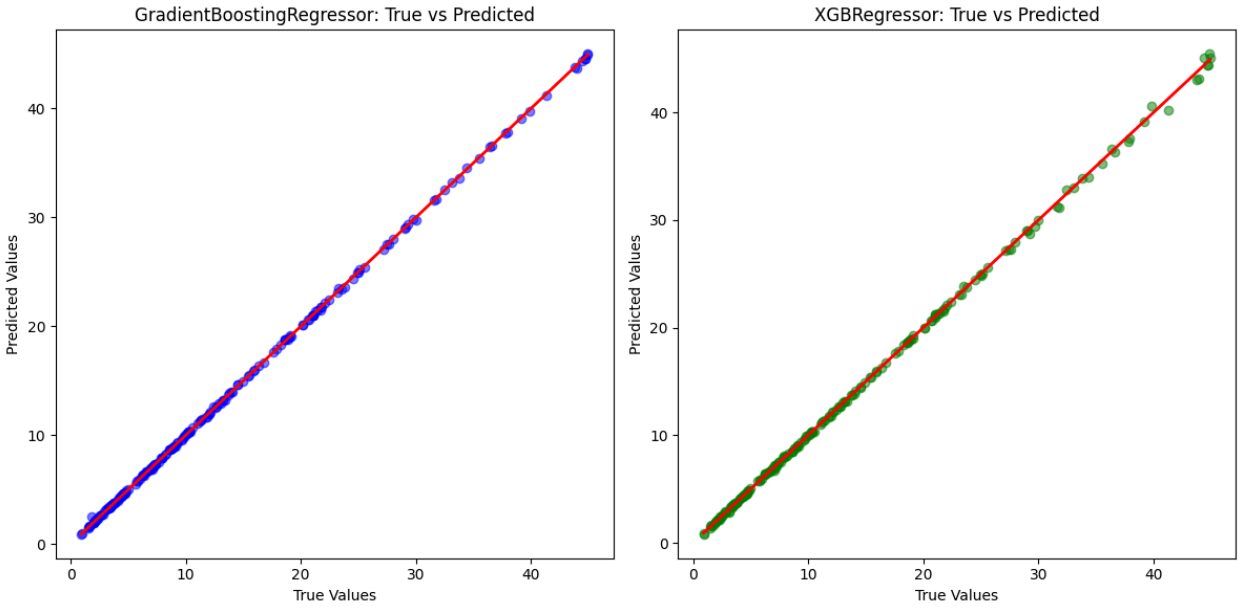

Fig-5: Actual Vs Predicted Both Model

Explanation:

GradientBoostingRegressor (left) and XGBRegressor (right). Both graphs show how closely the predicted values (on the y-axis) match the true values (on the x-axis). The red diagonal line represents perfect predictions, where the predicted values exactly equal the true values.

Looking at the plots, the dots for both models align almost perfectly with the red line, indicating excellent performance. However, to check for overfitting, we need to observe whether one model consistently predicts with less variation or shows any irregular patterns. In this case, the XGBRegressor (right) seems to perform slightly better as the green dots are more tightly clustered around the red line, showing fewer deviations. This indicates a very high fit to the data but raises the question of potential overfitting, as XGB models are often prone to fitting too closely on the training data.

The GradientBoostingRegressor (left) also performs well, with predictions closely matching the true values, but the blue dots show very slight scattering compared to XGB. This could suggest that the Gradient Boosting model generalizes slightly better.

In summary, both models perform well, but XGBRegressor might be slightly overfitting due to its extremely close alignment with the line, whereas GradientBoostingRegressor shows a balanced and robust fit.

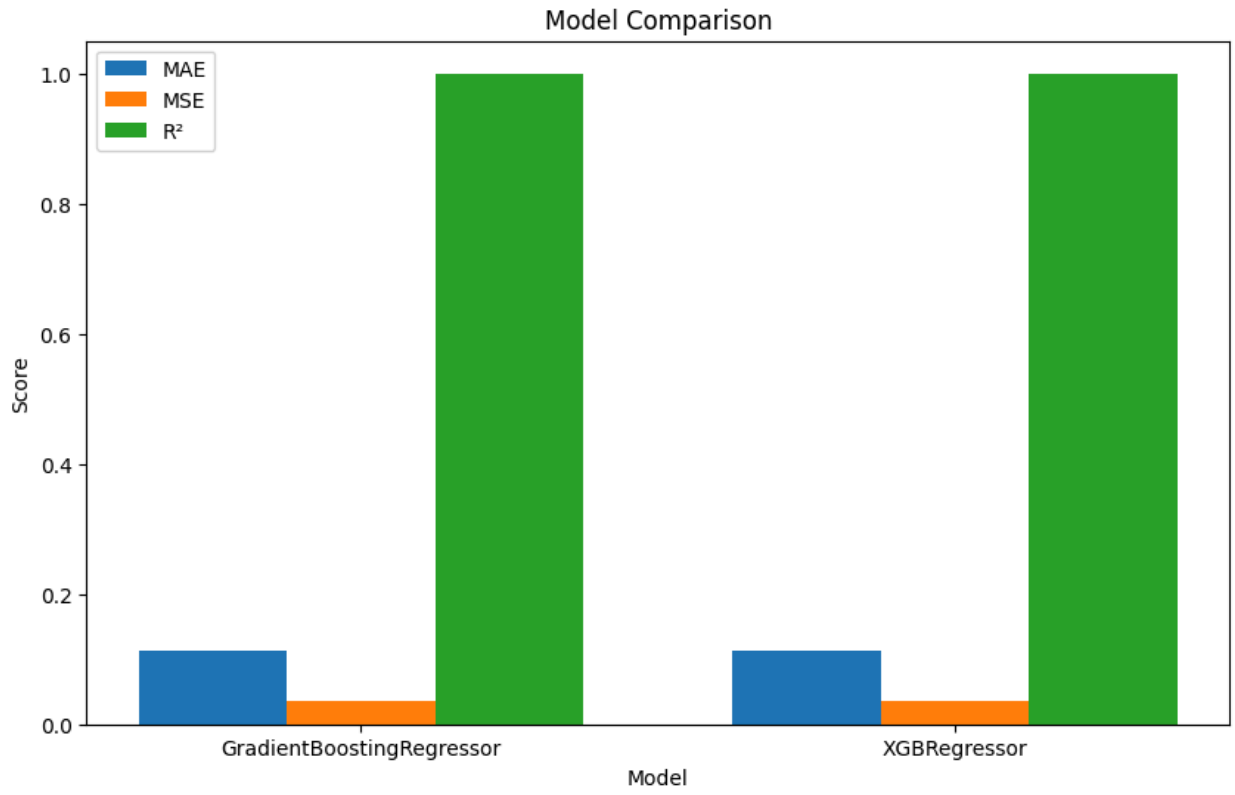

Fig-5: Model Comparison(MAE, MSE, R^2)

Explanation:

This bar chart compares the performance of two models, GradientBoostingRegressor and XGBRegressor, using three key metrics: MAE (Mean Absolute Error), MSE (Mean Squared Error), and R^2 (R-squared). The MAE values, shown in blue, are low and nearly identical for both models, indicating that the average prediction errors are minimal. Similarly, the MSE, represented by orange bars, is also very small, suggesting that both models handle larger errors well, as MSE penalizes large deviations more heavily.

The R^2 scores, shown as green bars, are extremely close to 1 for both models, meaning they explain nearly all the variance in the target variable. This high R^2 value indicates excellent performance, but it also raises a concern about possible overfitting, especially if the models are tested only on training or highly similar data. While both models perform almost equally in terms of these metrics, further evaluation on unseen data would help confirm their ability to generalize well. Overall, both GradientBoostingRegressor and XGBRegressor demonstrate outstanding predictive performance based on this comparison.