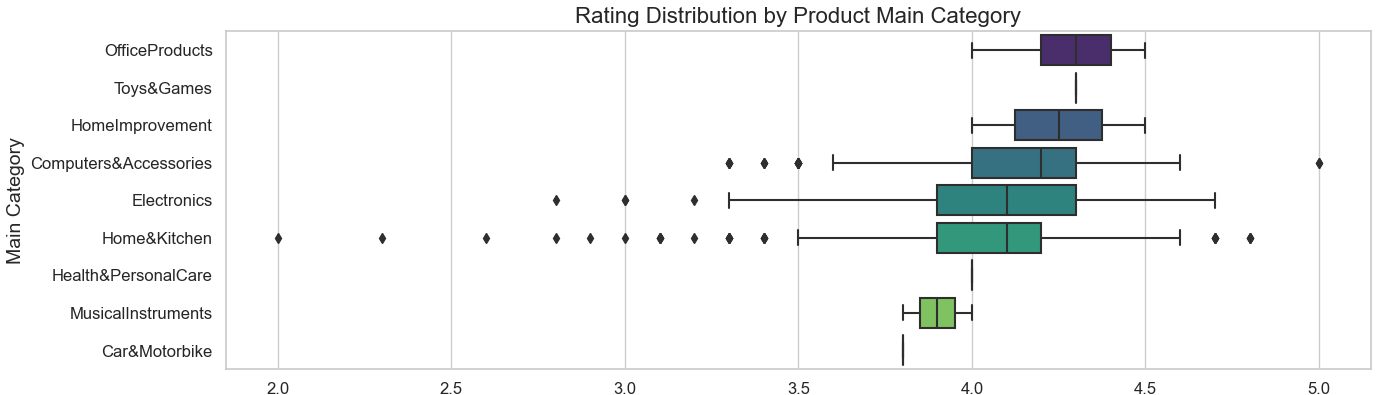
**Remark Data Community - Competency Test.**

Using tools of your choice complete the following tasks. Keep it simple. The goal of exercise is not to do the deepest EDA, build the best model possible, or spend lots of time building a robust solution. The goal is to demonstrate your ability to complete the end-to-end process in a reasonable timeframe and explain your thought process clearly.

1. Load data
   1. Load the data into the environment of your choice.
2. Clean data
   1. Clean the data so that it is more usable for subsequent steps. At least two columns should be transformed.
      1. I initially attempted to transform the discounted\_price, actual\_price, and discount\_percentage columns by removing currency symbols and converting them to numeric types.
      2. Missing values should be removed
3. Feature engineering
   1. Transform at least one column into another column or columns, which have potential use for exploratory data analysis and/or modeling.
      1. I created these two new futures:
         1. price\_reduction: Represents the difference between the actual price and the discounted price. This feature might provide insights into the amount of discount provided in absolute terms.
         2. discount\_to\_price\_ratio: Calculates the ratio of the discount amount to the actual price. This feature gives a relative sense of how significant the discount is compared to the actual price.
      2. The category column seems to contain hierarchical or combined data, with different levels separated by the pipe (|) symbol. This suggests that there are multiple layers of categorization for each product.
4. Exploratory Data Analysis (EDA)
   1. How complete is the data?
      1. **Data Quality and Completeness:** The dataset appeared to be quite complete with a low percentage of missing values. However, the quality and consistency of categorical data like category needed attention, especially for advanced analyses.
      2. **Potential for Predictive Modeling:** Variables like rating could be interesting targets for predictive modeling. The correlations and distributions of various features would inform feature selection and model design.
      3. **Business Insights:** From a business perspective, understanding these distributions and relationships can aid in pricing strategies, marketing, and inventory management.
      4. The dataset seems rich with information for both descriptive and predictive analyses. However, caution should be taken regarding potential biases (like in the distribution of ratings) and the depth of categorical data. Any modeling or further analysis should consider these factors to ensure accurate and meaningful results.
      5. Are there missing values?
         1. Given the small percentage of missing values, the data can be considered quite complete. However, the missing values will need to be addressed, either by imputation or by removing the affected rows, depending on the context and the importance of the missing data for the analysis.
            1. There are missing values in several columns, but the percentages are quite low, ranging from about 0.14% to 0.34%.
            2. product\_name: Approximately 0.14% missing values.
            3. category, discounted\_price, actual\_price, discount\_percentage, price\_reduction, and discount\_to\_price\_ratio: Each has about 0.20% missing values.
            4. rating: Around 0.27% missing.
            5. rating\_count: About 0.34% missing.
   2. Create visualizations which explain the distributions of ‘category’, ‘actual\_price’, ‘discount\_price’, ‘discount’, ‘rating’, and ‘rating\_count’
      1. **Category Distribution:** The category column, particularly the first level or two of the hierarchy, likely offers valuable insights into product groupings. The top categories could indicate popular or common product types in the dataset. However, deeper levels in the hierarchy might be too granular for some types of analysis.
      2. **Price-Related Features**:
         1. **Actual and Discounted Prices:** Both showed a wide range of values. The distributions might be right-skewed, indicating a concentration of products in the lower price range with fewer high-priced items.
      3. **Ratings and Rating Counts:**
         1. **Ratings:** The distribution seemed to center around higher ratings, suggesting that most products have good ratings. However, this could also indicate a bias in the data towards positively reviewed products.
         2. **Rating Counts:** There was likely a wide variance in the number of ratings per product, with many products having few ratings and a few products having a very high number of ratings.



A graph of a bar chart

Description automatically generated with medium confidence

* + 1. A screenshot of a graph

       Description automatically generated

1. If you have decided to engineer new columns based on any of the above columns, you may use the columns you created instead.
   1. **Price Reduction and Discount to Price Ratio:** These engineered features provide additional perspectives on the discounts. The distribution of price reductions and discount ratios could be helpful in understanding the extent of discounting practices.

A graph of a graph

Description automatically generated with medium confidence

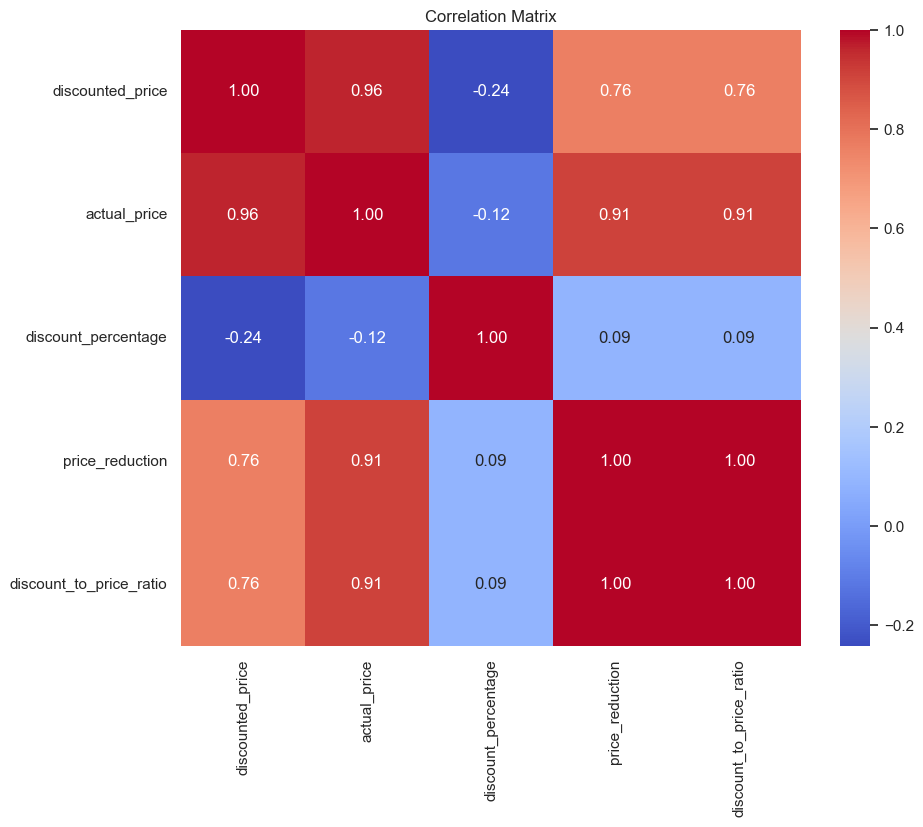
1. Are any of these variables correlated? **Please see the next page matrix.**
2. How do you know? The most common method to calculate this correlation is Pearson's correlation coefficient.

**1 indicates a perfect positive correlation.**

**-1 indicates a perfect negative correlation.**

**Values close to 0 suggest no linear correlation.**

**Color Coding:** The colors provide a visual aid to spot high or low correlations. Warmer colors (like red) indicate higher positive correlations, and cooler colors (like blue) indicate higher negative correlations.



**Correlations:**

If certain variables like actual\_price and discounted\_price showed strong correlations, it indicates a consistent discounting pattern.

Correlations involving rating and rating\_count could provide insights into consumer behavior, such as whether more expensive products tend to receive higher ratings or if higher discounts lead to more reviews.

1. Model
   1. Build a model to predict ‘rating’. In the interest of time, the accuracy does not need to be high.
      1. What modeling approach did you use and why? **3 different models (a Linear Regression model and also two Ensemble models)**
         1. **Linear Regression Modelling approach**
            1. Simplicity and Interpretability
            2. The target variable, rating, is continuous and numerical, which suits the regression analysis
            3. Linear Regression serves as a benchmark, and its performance on the dataset provides a basis for comparison with more sophisticated models if needed.
            4. It needs less computational power and less time for training and prediction compared to more complex models
            5. The initial assumption was that there might be a linear relationship between the features (like price, discount percentage, etc.) and the product ratings. Linear Regression is the go-to model to test this linear relationship.

A graph with blue dots and a red line

Description automatically generated

* + - 1. **Ensemble Techniques (Random Forest and Gradient Boosting Regression)**

**A graph of a number of trees

Description automatically generated** **A graph showing the number of trees

Description automatically generated**

* 1. What is the accuracy of your model?

**Random Forest MSE: 0.0667**

**Gradient Boosting Regression MSE: 0.0740**

**Linear Regression MSE: 0.0818**

* + 1. Why did you choose this accuracy metric?
       1. **MSE** is a standard metric for regression problems. It measures the average squared difference between the estimated values (predictions) and the actual value (true ratings), making it directly aligned with the objective of minimizing errors in a regression model.
          1. **Sensitivity to Large Errors**: MSE squares the errors before averaging, which means it disproportionately penalizes larger errors. This property is particularly useful when it's crucial to avoid large deviations from the true values, as might be the case in rating predictions.
          2. **Easy Interpretation Relative to the Target Variable**: MSE is in the same unit as the squared target variable, which, in this case, is the product rating. This makes it somewhat intuitive when assessing the model's performance relative to the scale of the ratings. Additionally, the square root of MSE, known as Root Mean Squared Error (RMSE), can be used to bring the metric back to the same scale as the target variable for even easier interpretation.
          3. **Baseline Comparison**: MSE provides a baseline against which the performance of more complex models can be compared. It's useful to know how much improvement (if any) is gained by employing more sophisticated models or techniques.
          4. **Simplicity and Common Usage**: MSE is straightforward to calculate and is widely used, making comparisons with other models or benchmarks easier and more meaningful.
       2. **Other metrics**:
          1. **MAE** provides a direct average of absolute errors, which can be more intuitive and less sensitive to outliers compared to MSE.
          2. **RMSE** brings the error metric back to the same scale as the target variable, making it more interpretable.
          3. **R-squared** measures the proportion of variance in the target variable that is predictable from the independent variables, offering insight into the model's explanatory power.

1. Understand and Explain
   1. What is the relative importance/significance of the features you used in your model?

To understand the relative importance or significance of features in a Linear Regression model, we typically look at the coefficients of the model. These coefficients indicate how much the dependent variable (rating, in this case) is expected to change when the feature changes by one unit, holding other features constant.

However, it's crucial to note that the scale of the features can significantly affect the magnitude of the coefficients. Features with larger scales can have smaller coefficients, and this does not necessarily mean they are less important. Therefore, to properly interpret the relative importance of features, it's often advisable to standardize the features before training the model.

Coefficient Absolute Coefficient

**discount\_percentage -2.437648e-03 2.437648e-03**

discount\_to\_price\_ratio -1.749038e-04 1.749038e-04

price\_reduction 1.226320e-04 1.226320e-04

discounted\_price -6.199519e-05 6.199519e-05

actual\_price 6.063685e-05 6.063685e-05

rating\_count 7.317334e-07 7.317334e-07

**discount\_percentage(Coefficient: -0.00244):**

This is the most influential feature based on its absolute coefficient.

The negative sign indicates an inverse relationship with the rating: as the discount percentage increases, the rating tends to decrease slightly.

* + 1. What importance/significance metric did you choose and why?

For the Linear Regression model, I chose to interpret the importance and significance of the features based on the coefficients of the model.

**Direct Interpretability:** In a Linear Regression model, each coefficient represents the change in the dependent variable (in this case, the rating) for a one-unit change in the corresponding feature, assuming all other features remain constant. This direct relationship offers clear interpretability of how each feature influences the target variable.

**Magnitude and Direction:** The magnitude of each coefficient indicates the strength of the impact, and the sign (positive or negative) indicates the direction of the relationship. This helps in understanding whether a feature positively or negatively affects the rating, and to what extent.

**Standard Practice:** Using coefficients as a measure of feature importance is standard practice in Linear Regression analysis. It is a straightforward and widely accepted method for assessing which features have the most significant influence on the prediction.

**Simplicity:** Given the context of demonstrating an end-to-end process within a reasonable timeframe, using coefficients as an importance metric aligns with the need for simplicity and efficiency.

**Relative Importance:** By comparing the absolute values of the coefficients, we can assess the relative importance of the features. This helps to identify which features are most impactful in the model.

* 1. Describe the relationship between the most important variable and the prediction.

The coefficient of discount\_percentage is negative, indicating an inverse relationship with the product rating.

This means that as the discount\_percentage increases, the product rating is expected to decrease slightly, assuming all other variables remain constant.

**Interpretation:**

The negative relationship suggests that higher discounts might be associated with slightly lower ratings.

**This could imply various things:**

Customers might perceive higher discounts as indicative of lower quality, affecting their ratings.

Products requiring higher discounts to sell might inherently have characteristics that lead to lower customer satisfaction.

It's also possible that the relationship is not directly causal but is influenced by other factors not captured in the model (e.g., the type of product, seasonal variations, customer expectations, etc.).

**Caution in Interpretation:**

While discount\_percentage shows the strongest relationship among the variables considered, its coefficient is still relatively small, suggesting that the effect on rating is modest.

Linear Regression coefficients represent an average effect across all observations, so individual variations can occur.

The context of the data and the business logic should always be considered when interpreting these results.

The interpretation is based on the assumption of a linear relationship and does not imply causation.

**In summary**, the discount\_percentage shows a slight negative correlation with product ratings in the Linear Regression model. However, the actual impact on the rating is relatively small, and the observed relationship should be contextualized within the broader scope of the data and external factors. It's important to explore further, possibly with different models or additional data, to understand the nuances of this relationship better.

* + 1. What method did you use to examine this relationship and why?

**Direct Interpretability in Linear Regression:** In a Linear Regression model, each coefficient quantitatively describes the expected change in the dependent variable (here, rating) for a one-unit change in the predictor (here, discount\_percentage), assuming all other predictors are held constant. This direct interpretation makes it straightforward to understand the nature and strength of the relationship.

**Simplicity and Efficiency:** Given the time constraints and the goal of demonstrating an end-to-end process, using the coefficients from a Linear Regression model is efficient and provides immediate insights without the need for complex computations or additional analysis tools.

**Quantitative Assessment:** The coefficient provides a numerical value that quantifies the relationship, making it easier to compare the impact of different features and understand their relative importance.

**Standard Analytical Practice:** Utilizing coefficients from Linear Regression is a standard practice in statistical analysis and data science for assessing relationships between variables. It’s a widely accepted method, especially when starting with exploratory analysis.

* 1. Do you believe that the ‘rating’ variable is able to be predicted based on the data available?
     1. Why or why not?

Based on the analysis and the results from the Linear Regression model and other models, the predictability of the rating variable based on the available data can be considered with some caveats:

**Reasons Why Rating Might Be Predictable:**

**Presence of Relevant Features:** The dataset includes features like actual\_price, discounted\_price, discount\_percentage, and rating\_count, which could logically influence a product's rating. For example, price and discounts might affect customer satisfaction and perception, thereby influencing ratings.

**Statistical Relationship Indicated:** The Linear Regression model, despite its simplicity, did show some level of association between these features and the rating. Although the coefficients were small, they suggest that there is at least some relationship.

**Scope for More Complex Models:** More sophisticated models (like Random Forests or Neural Networks) might be able to capture more complex relationships and interactions between features that a simple Linear Regression cannot, potentially improving predictability.

**Reasons Why Predicting Rating Might Be Challenging:**

**Subjectivity of Ratings:** Product ratings are inherently subjective and can be influenced by factors beyond what's captured in the dataset, such as individual customer preferences, expectations, or experiences.

**Small Coefficients in Linear Model:** The small magnitude of the coefficients in the Linear Regression model suggests that the features, as currently presented, might have a limited impact on the rating. This could imply that the model is missing critical factors or that the relationships are not strictly linear.

**Potential for Non-Linear Relationships:** The relationships between features and ratings might be non-linear or influenced by unobserved variables, which Linear Regression cannot adequately capture.

**Need for More Granular Data:** More detailed data, such as text reviews, customer demographics, or more granular product categories, might be necessary to build a more accurate predictive model for ratings.

**Conclusion:**

While there is potential to predict the rating based on the available data, the effectiveness of this prediction is likely to be limited by the current feature set and the linear nature of the initial model. There is potential for improvement with more complex modeling techniques and additional, perhaps more qualitative, data. However, it's also important to acknowledge the inherent challenges in predicting subjective outcomes like product ratings.

* 1. Suggest at least two ways to improve the model performance.
     1. **Feature Engineering and Selection:**
        1. **Incorporate More Features:** If additional data is available, consider incorporating more features that could influence product ratings. This might include product specifications, brand information, time-related features (like seasonality or time since product launch), customer demographics, or even aggregate measures derived from customer reviews (like sentiment scores).
        2. **Text Analysis of Reviews:** If textual customer reviews are available, performing natural language processing (NLP) to extract sentiment, topics, or key phrases could provide valuable predictive insights.
        3. **Interaction Terms:** Consider adding interaction terms to your model where appropriate. For instance, the interaction between discount\_percentage and actual\_price might have a different effect on ratings than either feature alone.
        4. **Polynomial Features:** Explore polynomial or non-linear transformations of features. This is particularly useful if the relationship between the features and the target variable is not linear.
     2. **Try More Complex Modeling Techniques:**
        1. **Ensemble Methods:** Algorithms like Random Forest or Gradient Boosting Machines (GBM) can capture complex patterns in the data better than a simple Linear Regression. They are particularly good at handling non-linear relationships and interactions between features.
        2. **Regularization Techniques:** If overfitting is a concern (especially when using a large number of features), consider using regularization techniques like Ridge or Lasso Regression. These methods add a penalty to the model for having too many large coefficients, which can help in reducing overfitting.
        3. **Neural Networks:** For large and complex datasets, particularly those with high-dimensional features (like text data), neural networks can be effective. They require more data and computational resources but can model complex, non-linear relationships.
        4. **Hyperparameter Tuning:** For models with hyperparameters (like Random Forest, GBM, or neural networks), use techniques like Grid Search or Random Search to find the optimal set of hyperparameters.
     3. **Additional Considerations:**
        1. **Data Quality**: Ensure the data is clean and well-preprocessed. Handling missing values appropriately, removing outliers, or normalizing features can significantly impact model performance.
        2. **Cross-Validation:** Use cross-validation techniques to assess the generalizability of your model. This helps in understanding how well the model performs on unseen data.
        3. **Model Evaluation Metrics:** Use appropriate and multiple evaluation metrics to assess model performance. Besides MSE, consider using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.
     4. Each of these strategies comes with its own set of considerations and trade-offs, such as model complexity, interpretability, and computational efficiency. The best approach often depends on the specific characteristics of the dataset and the business context of the problem.
        1. The Mean Squared Error (MSE) values I have obtained for the three models – Random Forest, Gradient Boosting, and Linear Regression – provide valuable insights into their performance in predicting product ratings. Here's a summary and interpretation:

**Random Forest MSE: 0.0667**

This is the lowest MSE among the three models, indicating that the Random Forest model has the best performance in terms of accuracy.

The Random Forest, an ensemble method that builds multiple decision trees and combines their results, seems to capture the complexities and non-linear relationships in your data better than the other models.

**Gradient Boosting Regression MSE: 0.0740**

The Gradient Boosting model also performs better than the Linear Regression but not as well as the Random Forest.

Like Random Forest, Gradient Boosting is an ensemble method that builds trees sequentially, each one correcting the errors of the previous. Its slightly higher MSE compared to Random Forest could be due to various factors like model parameters, the nature of the data, or how well it handles overfitting.

**Linear Regression MSE: 0.0818**

This model has the highest MSE, suggesting it's the least accurate for this particular task.

Given that Linear Regression assumes a linear relationship between the features and the target, this result could indicate that the relationships in your data are more complex than what a simple linear model can capture.

**Conclusions:**

**Model Complexity vs. Performance:** The improved performance of Random Forest and Gradient Boosting over Linear Regression highlights the potential complexity and non-linearity in your data. These ensemble methods are better suited for capturing complex patterns and interactions between features.

**Trade-off Between Simplicity and Accuracy:** While Random Forest provides the best accuracy, it's more complex and less interpretable than Linear Regression. This is a common trade-off in machine learning – more complex models often provide better accuracy but at the cost of interpretability and simplicity.

**Potential for Further Tuning:** There's room for further improvement, especially for the ensemble models. Hyperparameter tuning, feature engineering, and trying different model configurations could lead to even better results.

These insights should guide how you approach further modeling and analysis. If interpretability is crucial for your task, you might prefer Linear Regression despite its lower accuracy. If accuracy is more important, then focusing on improving and understanding the ensemble models would be the way forward.

* 1. In a few paragraphs or less, explain to a non-technical audience what you learned during this analysis. Keep your description as simple and easy to understand as possible.

During this analysis, I worked on predicting the ratings of products based on various features like their prices, discounts, and the number of people who have rated them. Think of it like trying to guess how much people will like a product based on its sale details.

**Here's what I found:**

**The Influence of Discounts:** Interestingly, products with bigger discounts don't always have higher ratings. It seems like if a product is on a big sale, people might not rate it as highly. This could be because people might think a heavily discounted product is not as good in quality.

**Price Isn’t Everything:** The actual price of a product and the price after discount don't have a big impact on how people rate the product. This means just because something is expensive or sold at a big discount doesn’t necessarily mean people will love or hate it.

**Every Little Bit Counts, But Not Too Much:** The features we looked at, like the price and discount, do affect the ratings a bit, but not as much as one might think. It's more about the overall experience with the product than just its price or discount.

**Predicting Ratings is Tricky:** Ratings are opinions, and everyone’s different. So, predicting how people rate a product based on numbers like price and discount isn't very straightforward. It's like trying to guess someone's favorite ice cream flavor just by knowing their age – there’s a lot more to it.

To improve our guesses in the future, we could look at more information. Maybe we can analyze what people say in their reviews or consider more details about the product. We also tried a simple way of predicting these ratings, but there are more complex methods that might do a better job, though they are harder to understand and use.

In short, we learned that the story behind product ratings is more than just numbers – it's about personal preferences, quality perceptions, and many other factors that aren't just about the price tag or the discount on it.