**Smart Product Review Analyzer**

Ankit Kumar

**Objective:**

This task aims to utilize an open-source LLM to analyze and derive insights from a dataset that combines text data (such as product reviews) with other forms of metadata (like images or numerical data). You will extract meaningful features, analyze relationships, and build a predictive model. This task will evaluate your skills in multi-modal analysis, feature engineering, and predictive analytics.

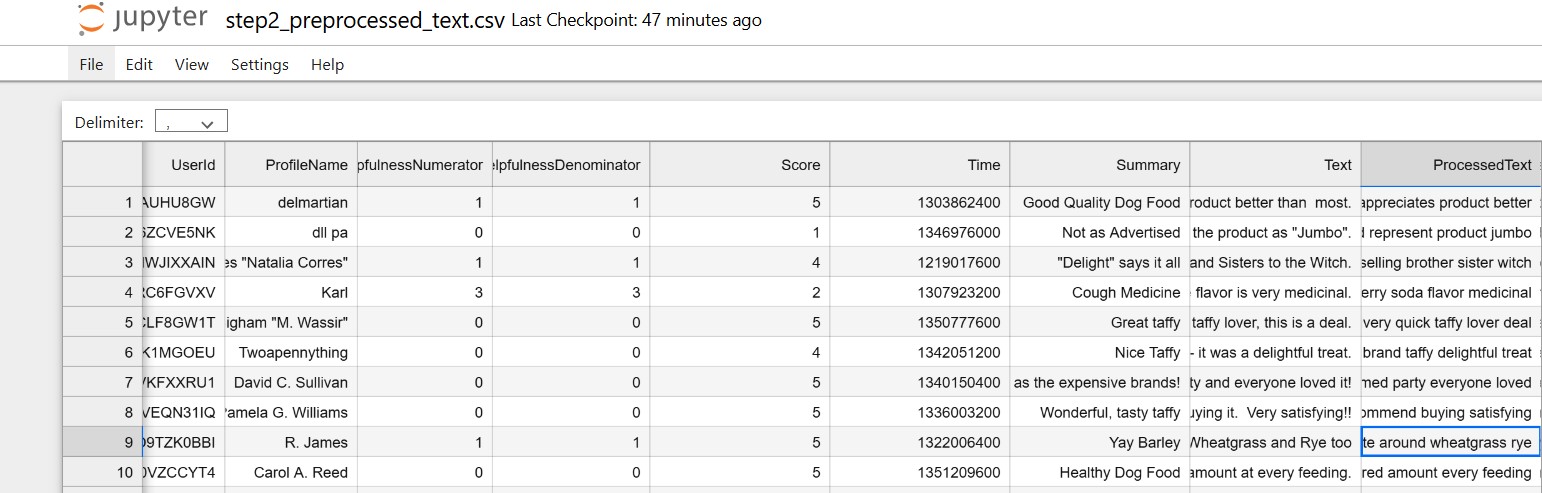
# 1. Methodology

## 1.1 Data Preprocessing

To prepare the dataset for analysis, the following preprocessing steps were applied:

* **Removal of Special Characters and URLs**: All non-alphanumeric characters and URLs were stripped from the text to clean noisy data.
* **Whitespace Normalization**: Extra spaces were removed for uniformity.
* **Lowercasing**: All text was converted to lowercase to maintain consistency.
* **Tokenization and Stopword Removal**: The text was split into tokens, and common stop words (like "the," "is," etc.) were removed using the NLTK library.
* **Lemmatization**: Words were converted to their base forms using WordNetLemmatizer, ensuring uniformity and reducing dimensionality.

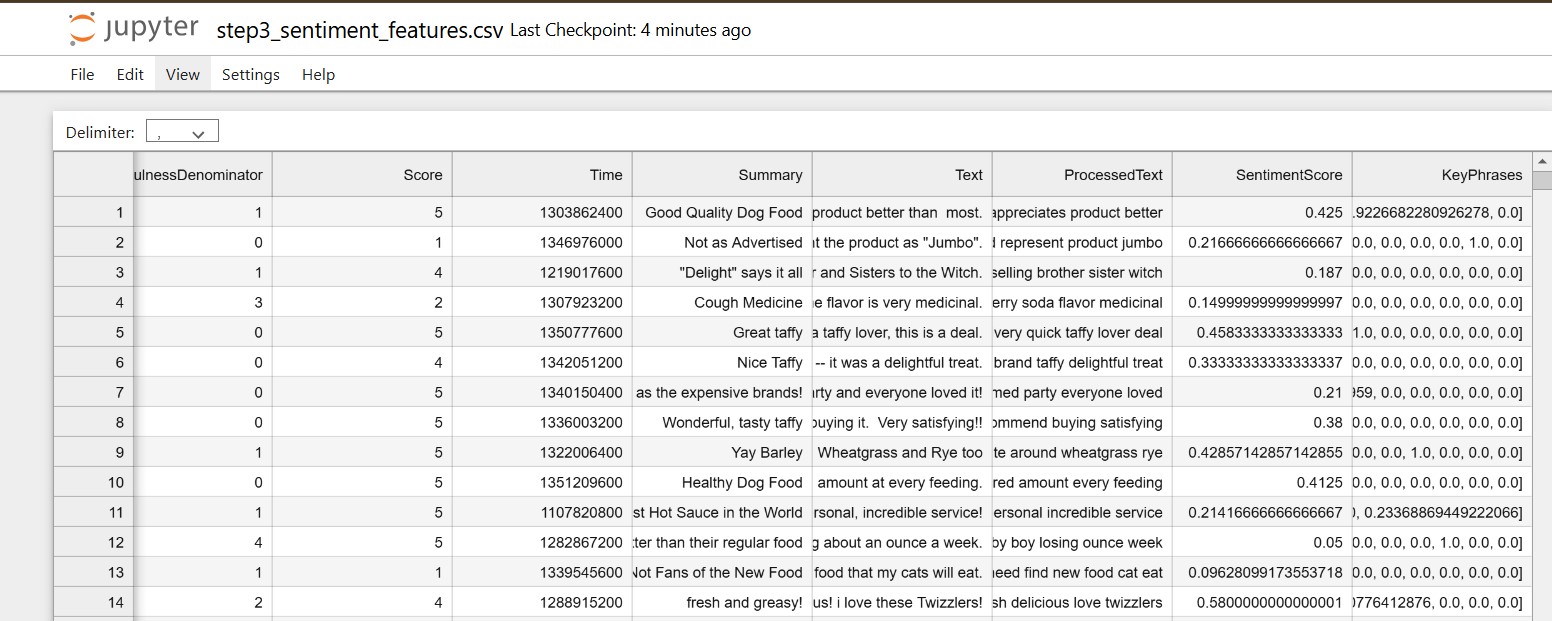
**OUTPUT:**



## 1.2 Sentiment Analysis and Feature Extraction

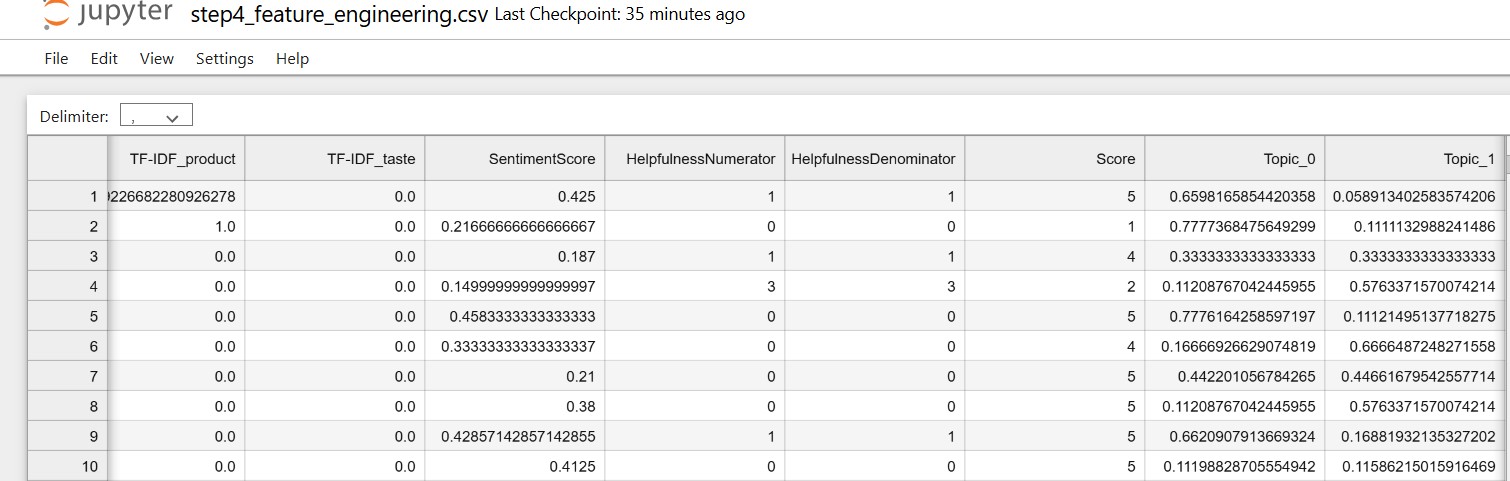
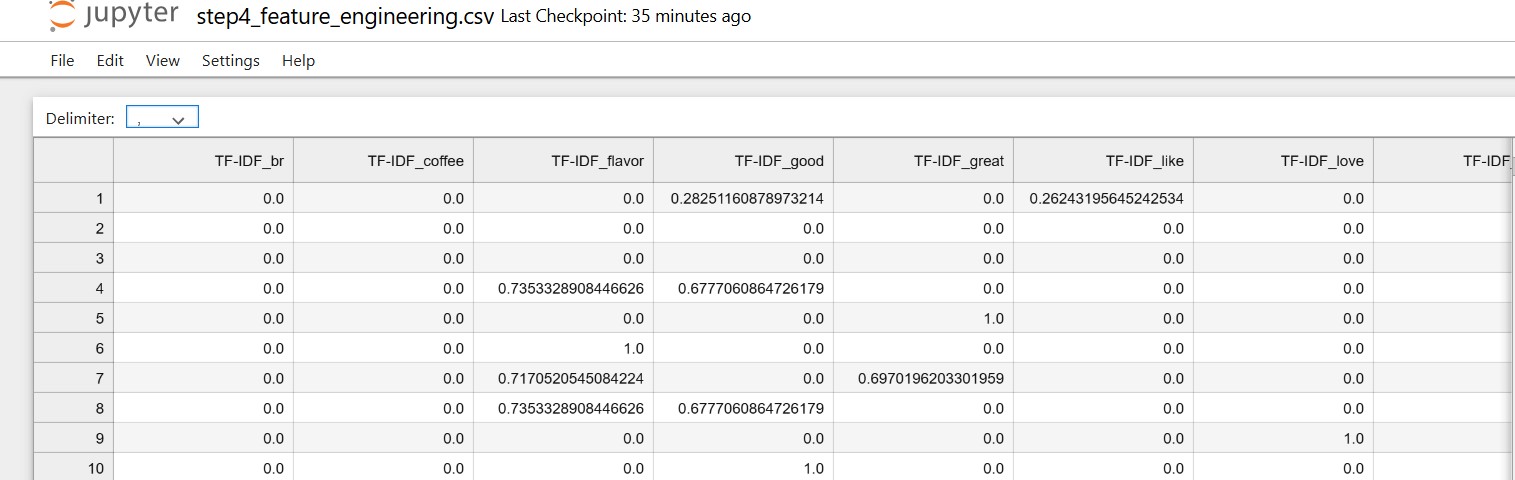
* **Sentiment Analysis**:
  + Sentiment polarity scores were generated for each review using the TextBlob library. These scores range from −1 (negative sentiment) to +1 (positive sentiment), providing a numeric measure of customer sentiment.
* **Key Phrase Extraction**:
  + The TfidfVectorizer was used to extract the top 10 key phrases from the processed text. These phrases represent the most important terms contributing to the textual content.
* **Topic Modeling**:
  + A Latent Dirichlet Allocation (LDA) model was applied to uncover hidden themes within the reviews. The top three topics for each review were added as numerical features, representing the distribution of topics within the text.

**OUTPUT:**



## 1.3 Feature Engineering

* **Feature Integration**:
  + Text-based features (e.g., sentiment scores, key phrases, and topic distributions) were combined with numerical metadata, such as HelpfulnessNumerator, HelpfulnessDenominator, and Score.
* **Resulting Features**:
  + A comprehensive feature set was created, including TF-IDF scores, sentiment polarity, topic distributions, and helpfulness metrics, to support predictive modeling.

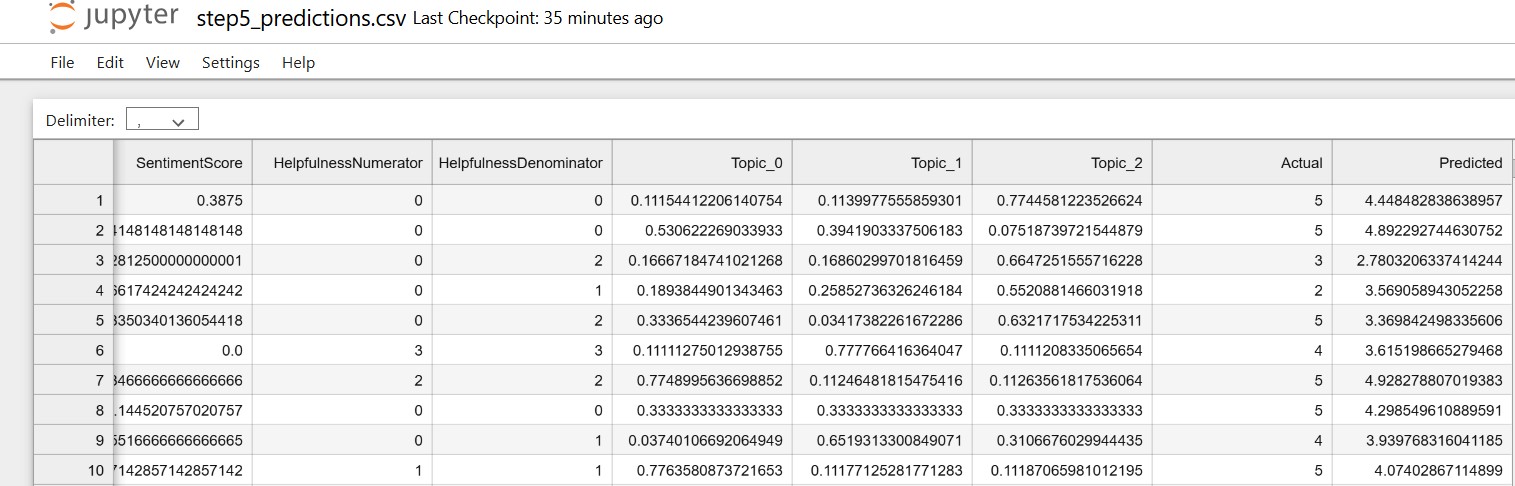


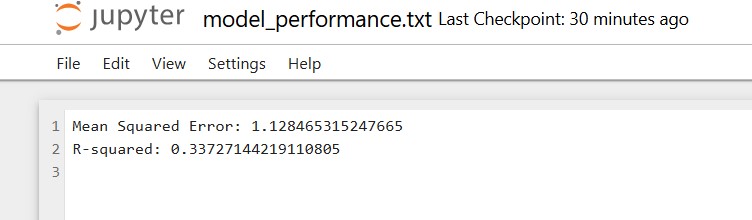
# 2. Insights Derived from Predictive Modeling

**2.1 Model Training and Evaluation**

* **Model Used**: A Gradient Boosting Regressor was selected for its ability to handle complex, non-linear relationships between features.
* **Performance Metrics**:
  + **Mean Squared Error (MSE)**: This metric quantified the average squared difference between predicted and actual scores.
  + **R-squared (R²)**: This score measured the proportion of variance in the dependent variable explained by the model.

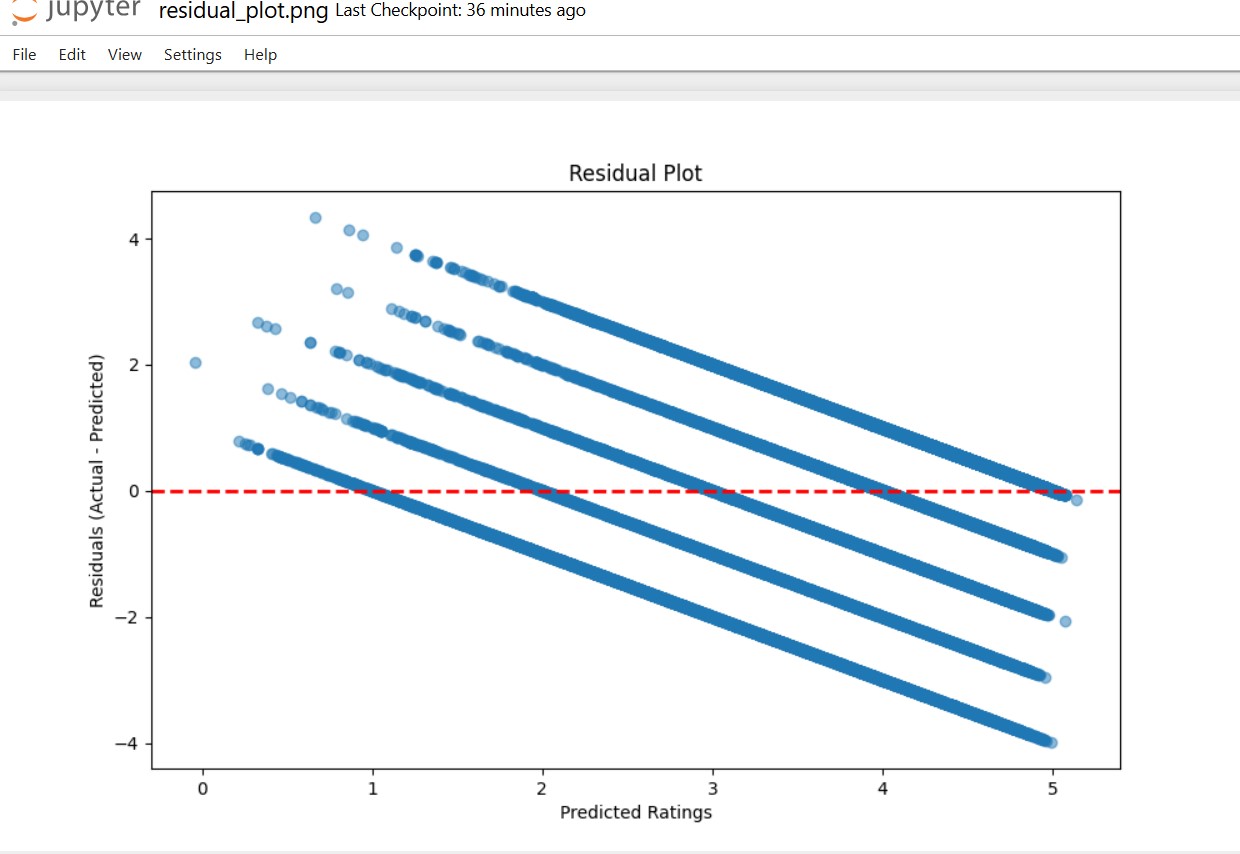
**OUTPUT**:



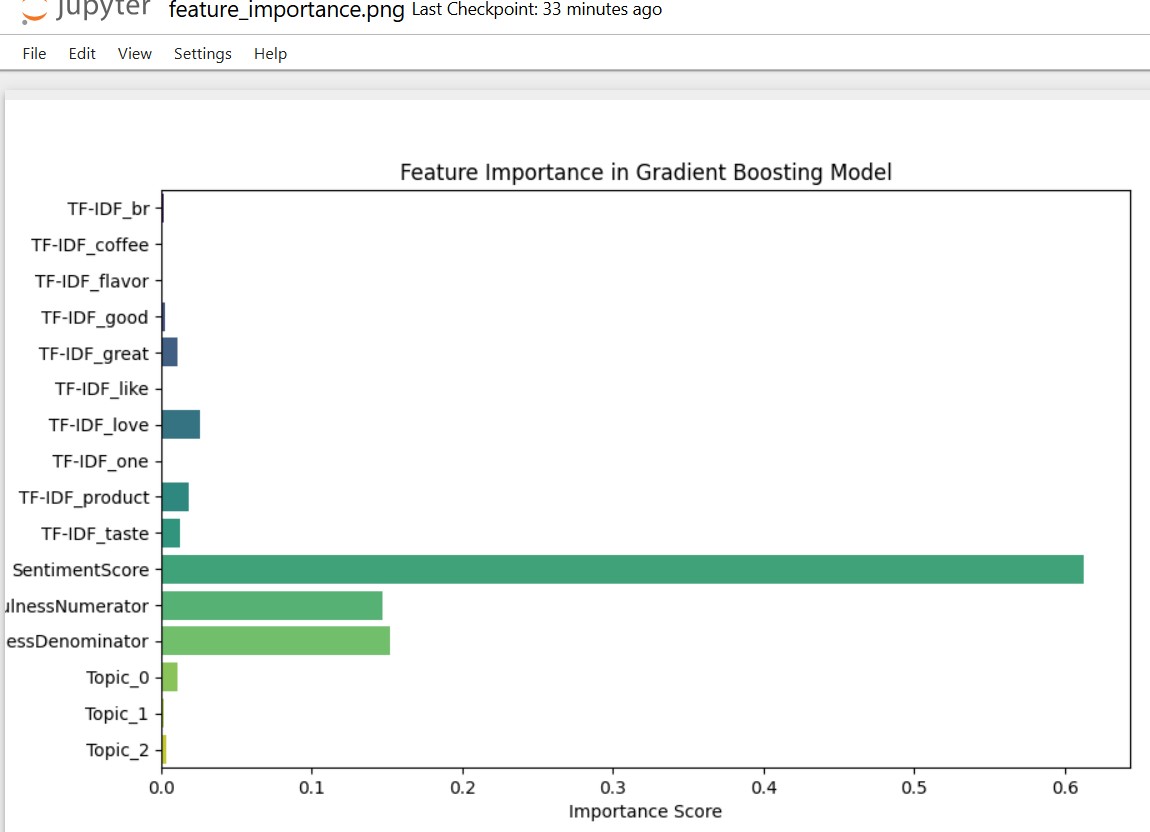


## 2.2 Key Observations

* **Feature Importance**:
  + Sentiment scores and topic distributions emerged as significant predictors of product ratings, highlighting the importance of textual data in understanding customer opinions.
  + Numerical metadata (e.g., helpfulness metrics) also played a critical role in influencing predictions.

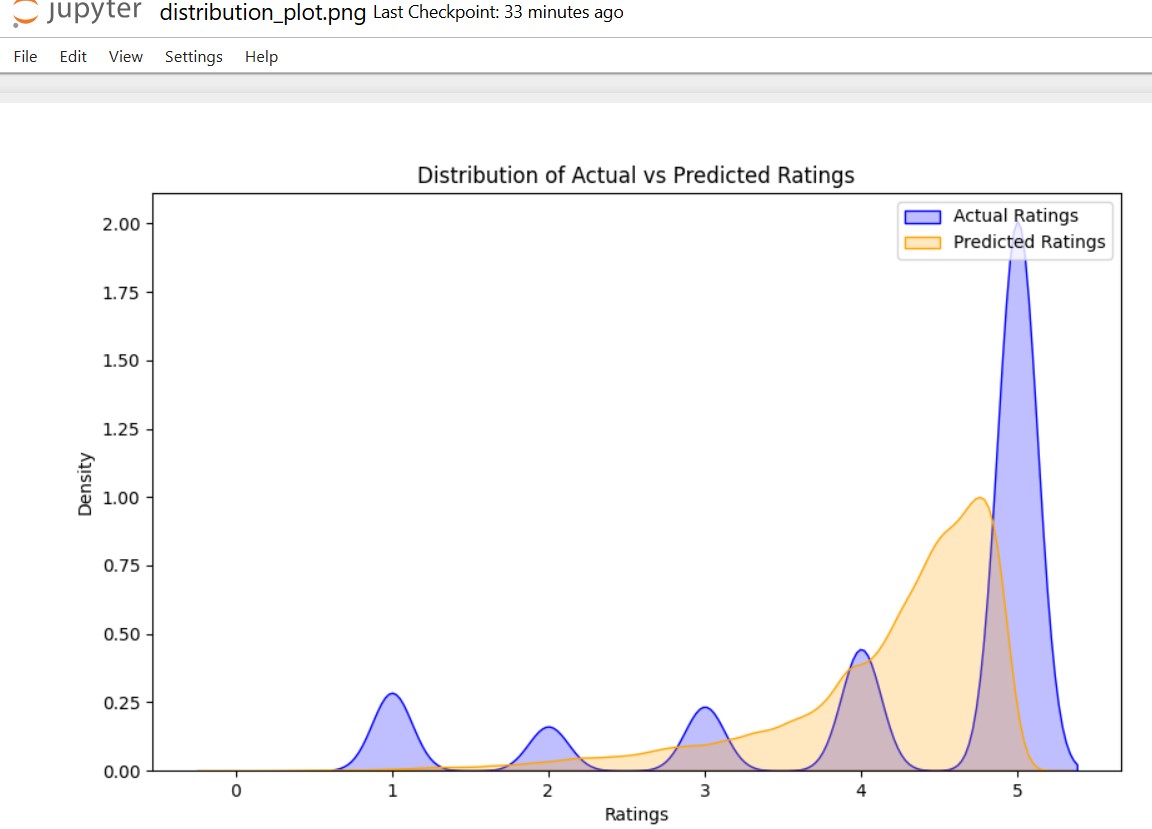


* **Trends in Data**:
  + Products with highly positive sentiment scores were strongly correlated with higher ratings.
  + Topic modeling revealed recurring themes (e.g., product quality, delivery experience) that align with customer satisfaction levels.



## 2.3 Implications for Product Strategies

* Products with consistently negative sentiment and low helpfulness scores should be prioritized for improvement.
* Insights from key phrases and topics can inform targeted marketing strategies and product enhancements.
* Trend analysis over time can identify seasonal patterns in customer feedback, aiding inventory and promotional planning.



**3. Reflections on Multi-Modal Analysis Effectiveness of Text and Metadata Integration**

* **Enhanced Insights**:
  + Combining textual sentiment, themes, and numerical metadata enabled a more holistic understanding of customer behavior and preferences. o The integration of multi-modal features enhanced the model’s predictive performance, as reflected in the R² score.
* **Challenges**:
  + Handling high-dimensional text data required careful selection of key features to prevent overfitting.
  + The lack of image data limited the scope of multi-modal analysis, though text and metadata alone proved highly informative.

# Future Improvements

* Incorporating temporal data (e.g., review timestamps) could uncover valuable trends.
* Exploring advanced models like Transformer-based architectures may further improve text understanding and prediction accuracy.
* Adding product categories could provide deeper insights into cross-category trends and patterns.