Group I

Elevating Sephora's Product Experience with an Innovative Recommender System.

Presented By:

- Dipti Baral
- Koshish Aryal
- Punam Bhattarai
- Shweta Laljibhai Thummar
- Uttam Tamang

Overview

What is our Aim and Objectives?

Sephora's vast selection of skincare products can overwhelm customers, making it challenging to find the right items. Using a dataset of Sephora products and skincare reviews, we aim to develop a recommendation system with collaborative and hybrid filtering techniques. This system will analyze customer reviews and preferences to deliver personalized skincare product suggestions, enhancing customer satisfaction and boosting sales.

Dataset

Understanding Dataset.

This dataset includes details on over 8,000 Sephora beauty products and 1 million user reviews for 2,000 skincare items. It covers product names, prices, ingredients, and user ratings, providing a rich source for analyzing skincare product performance and customer preferences.

KAGGLE

Source

154 MB

Size

1100554 / 40

Rows / Columns

Data Cleaning

Refining Data Quality with Data Cleaning.

Missing Values

Mean imputation was used for numerical columns and mode imputation for categorical columns to fill in the missing values in the columns.

Outliers

By applying the IQR approach, outliers in the Price column were found. A value was considered an outlier and eliminated if it was more than 1.5 times the IQR of the lower and upper quartiles.

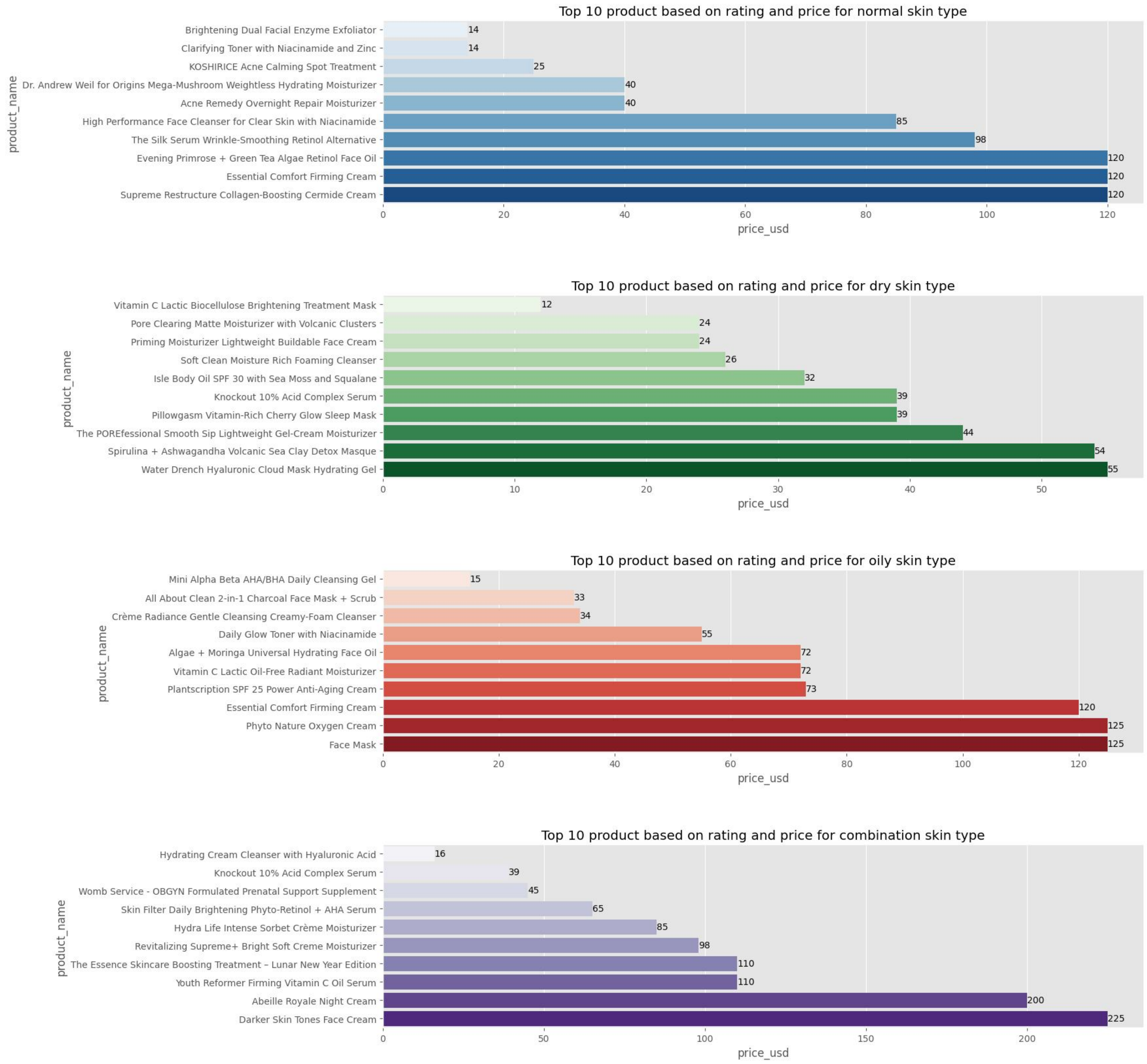
Duplicates

Removed redundant entries to preserve the accuracy of the data.

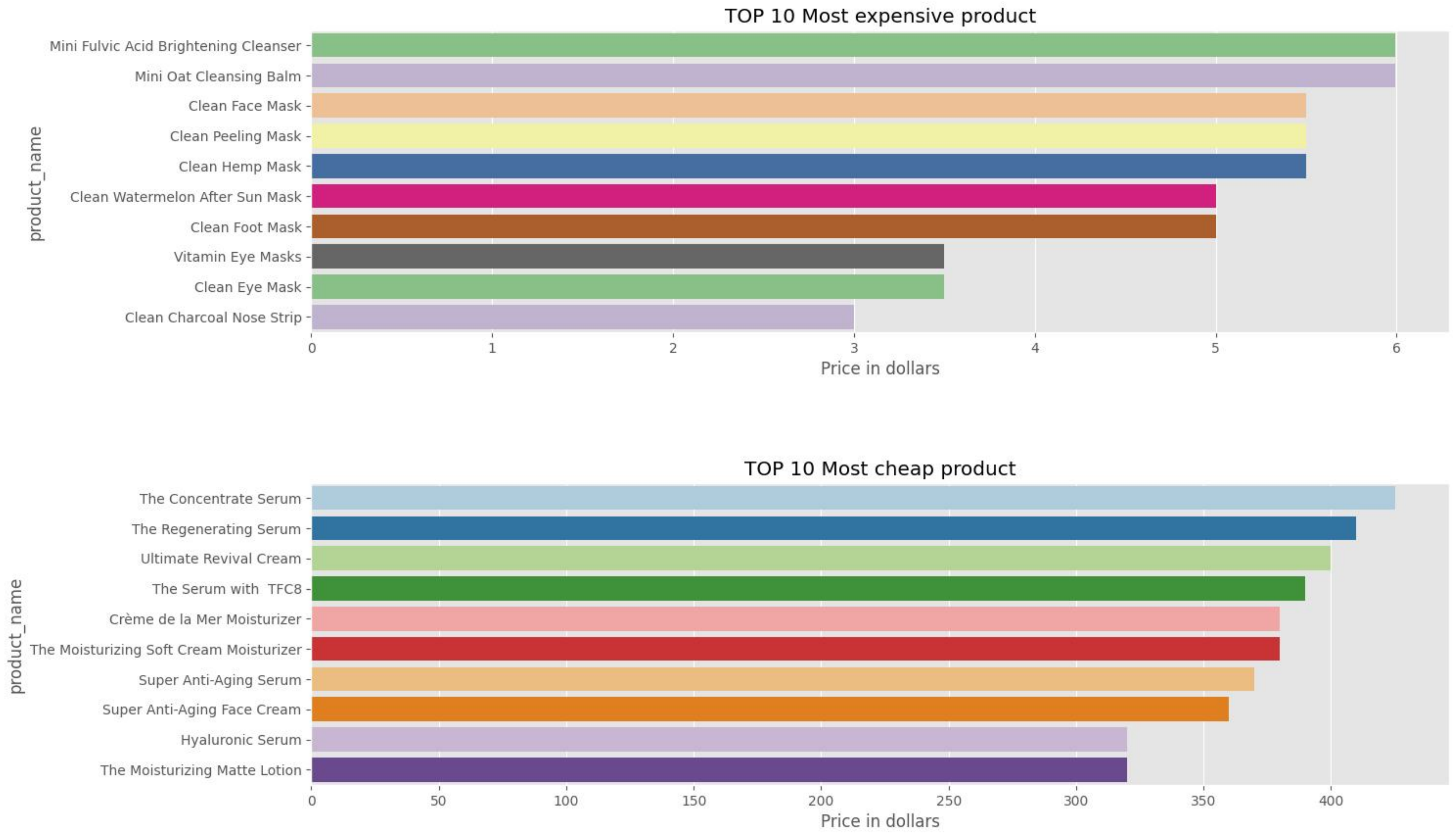
Data Visualization

Visualizing Key Insights

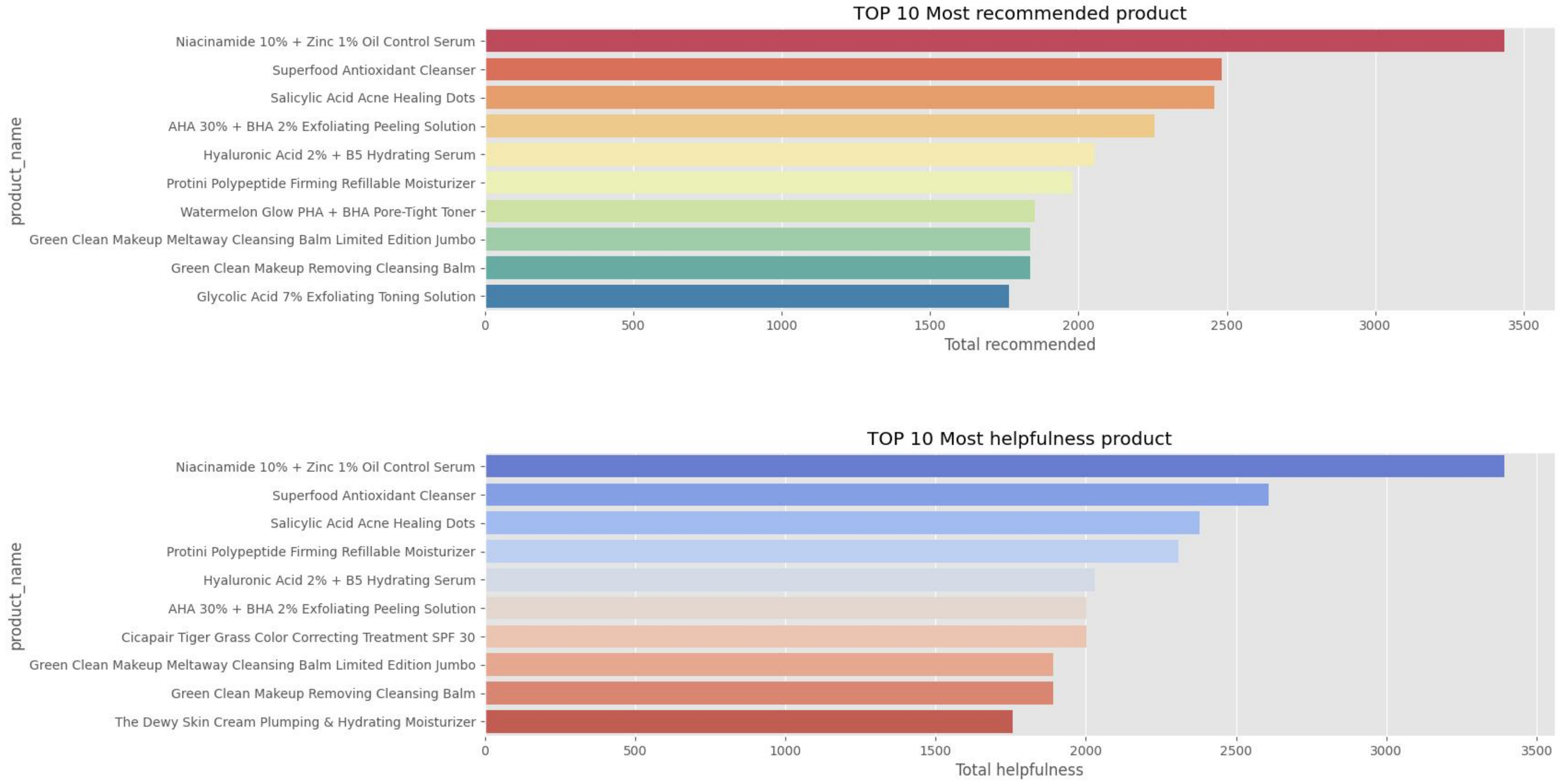
Most reccomended product for each skin types

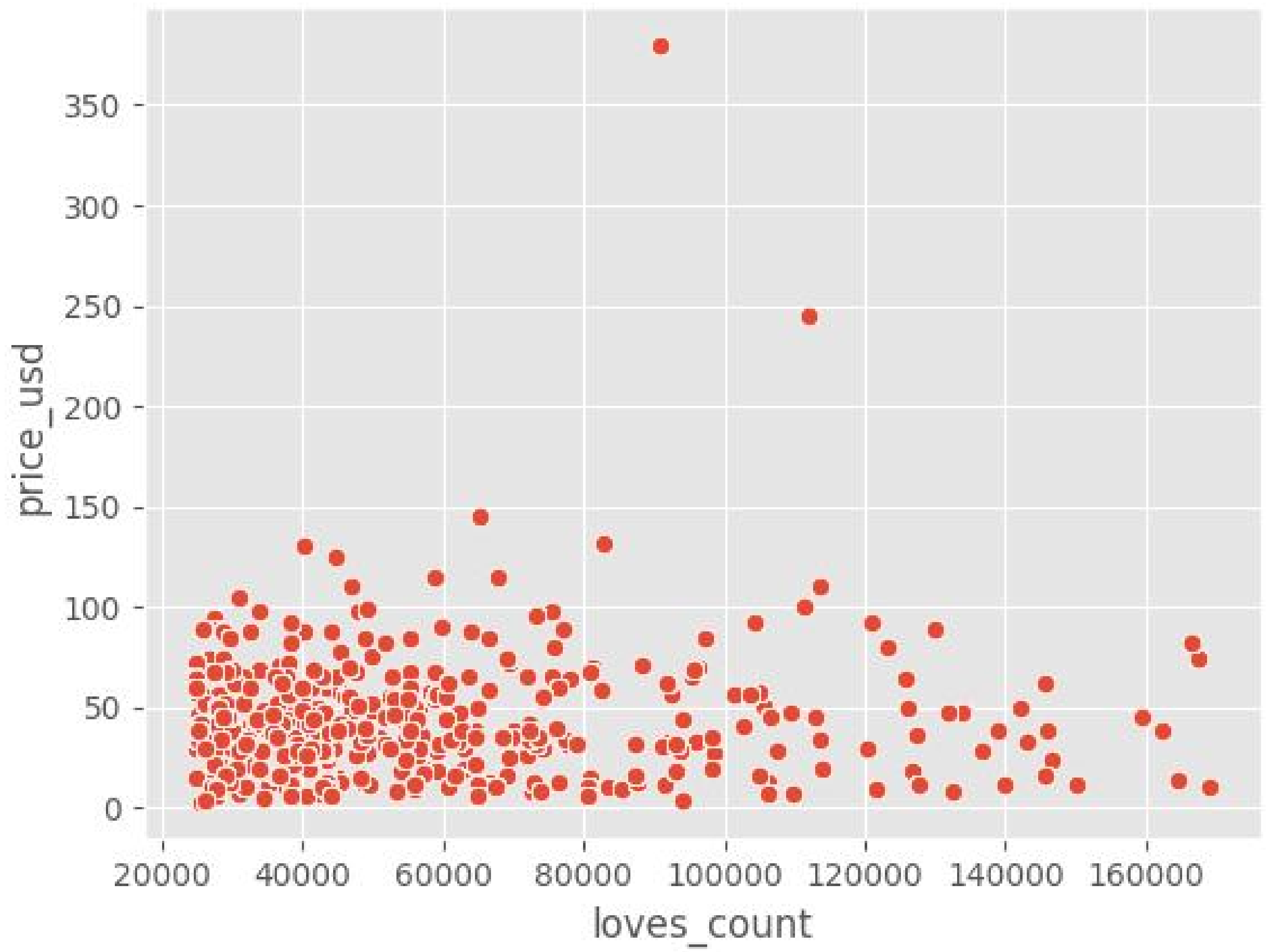


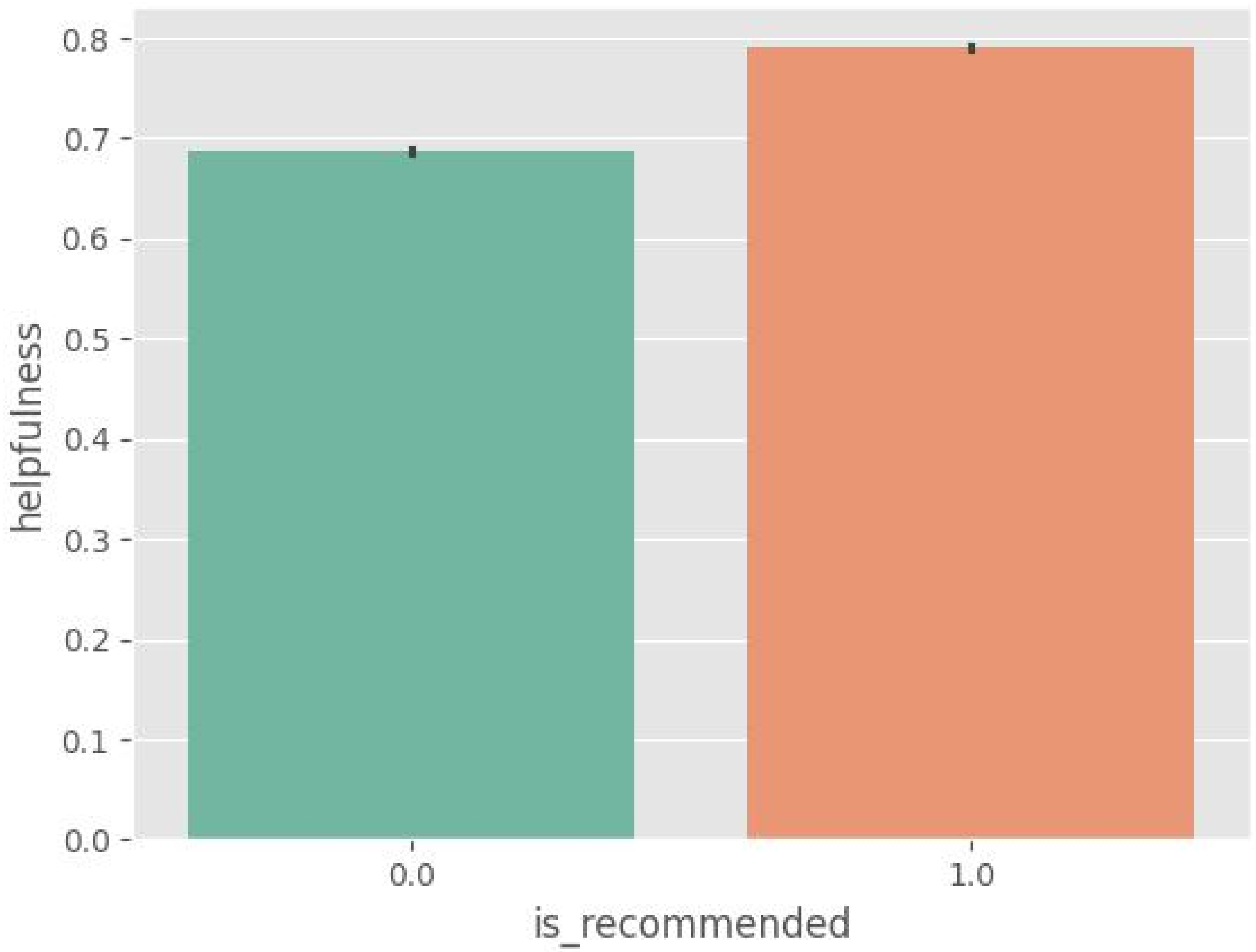
Top 10 Product based on price

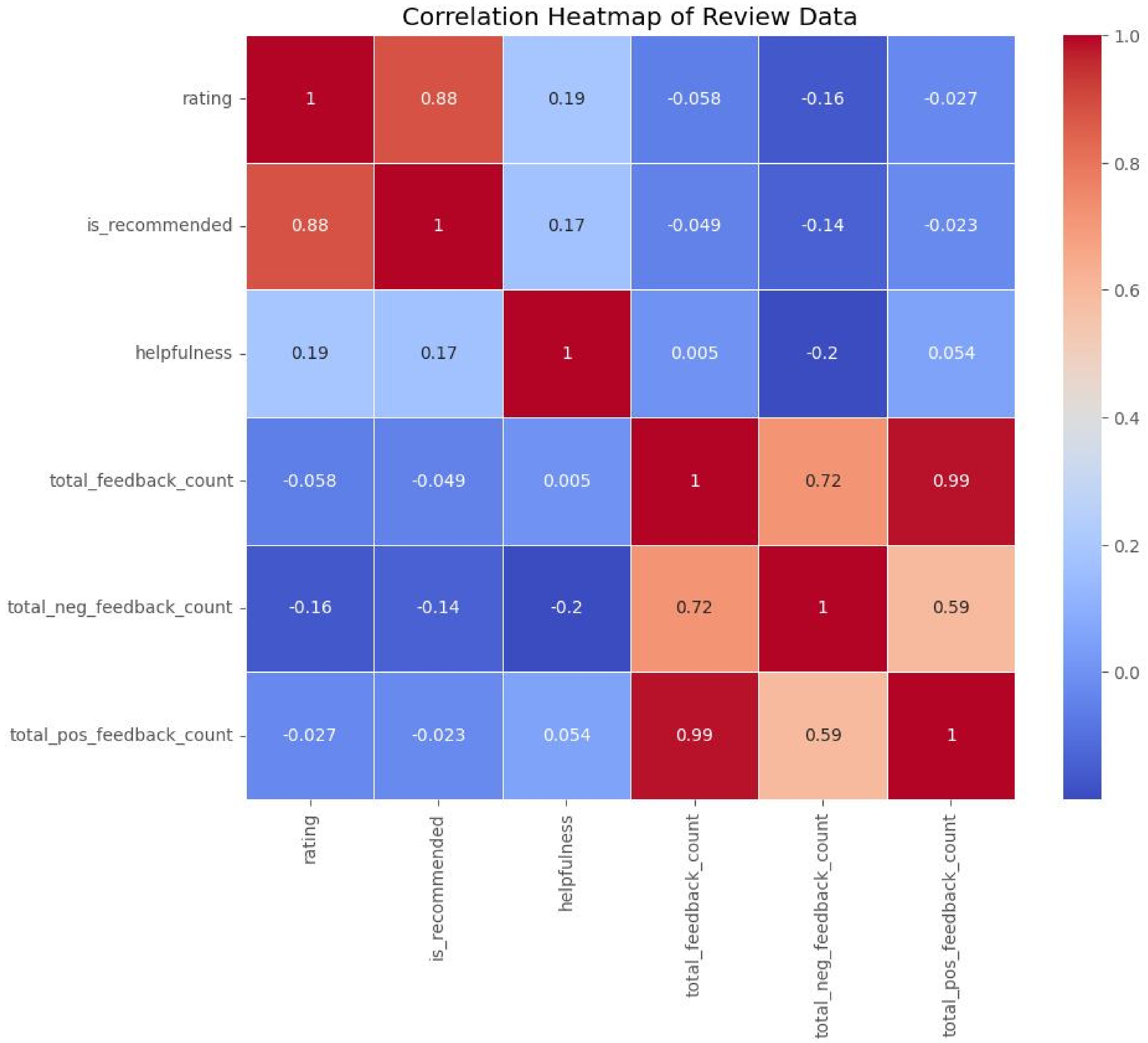


Top 10 Product









Heatmap (Correlation Matrix)

Rating & Recommendation: There's a strong positive correlation (0.88).
Feedback Counts: \`total_feedback_count\` is highly correlated with \`total_pos_feedback_count\` (0.99) and moderately with \`total_neg_feedback_count\` (0.72).
Helpfulness: Shows low correlation with other features, indicating it is largely independent.

Data Visualization

Interactive PowerBI Dashboard.

Methodology

Content-Based Filtering and Collaborative Filtering.

Content-based filtering recommends Sephora skincare products by analyzing their features, such as ingredients, brand, and user ratings. It suggests products similar to those a user has previously liked, tailoring recommendations to individual preferences.

Collaborative filtering focuses on user behavior, identifying patterns in how customers rate and purchase products. It recommends products based on what similar users have liked, even if the items don't share the same features.

Methodology

Hybrid Approach

Combining both methods, our hybrid model enhances recommendations by leveraging product features and user behavior.

This ensures Sephora customers receive personalized, accurate product suggestions, improving their shopping experience.

Methodology

Deep Learning Model.

Deep neural network created using **Keras**, wrapped with **KerasRegressor.**

Dense Layer

Fully connected layers with **128** and **64** neurons.

Dropout Layer

Dropout regularization to prevent **overfitting**.

Activation Functions

ReLU and **softmax** activation functions.

Optimizer

Adam and **RMSprop** optimizers.

Methodology

Large Language Model.

Our project integrates traditional machine learning techniques with **large language models (LLMs)** to enhance product recommendations. It employs Truncated SVD for dimensionality reduction, combined with sentiment analysis using the **SentimentIntensityAnalyzer** from NLTK.

Additionally, it leverages **BERT**, a transformer-based LLM, to capture deep contextual relationships within product reviews, refining the recommendation process through sophisticated feature extraction and classification pipelines.

Analysis & Result

Evaluation of Deep Learning Model.

Test MSE	**0.2094**
Test MAE	**0.3385**
Accuracy	**0.7615**

Analysis & Result

Evaluation of SVD with LLM.

	Precision	Recall	F1-Score	Support
Positive	0.58	0.38	0.46	10996
Neutral	0.00	0.00	0.00	0
Negative	0.88	0.91	0.89	46048
Accuracy			0.81	57044
Macro AVG	0.49	0.43	0.45	57044
Weighted AVG	0.82	0.81	0.81	57044

Discussion

Interpretation of results & Implications.

The best parameters identify the ideal batch size, epochs, optimizer, activation function, and dropout rate, highlighting the most effective settings for the dataset and model architecture.

Best Model Parameters

Model Optimizer	**Adam**
Model Dropout Rate	**0.3**
Model Activation	**Softmax**
Epochs	**200**
Batch Size	**32**

Discussion

Strength of the Project.

Hyperparameter Optimization

RandomizedSearchCV ensures that the model is tuned to its best performance by exploring a wide range of hyperparameter combinations.

Flexibility

The model can adapt to different configurations, making it versatile for various types of data and tasks.

Regularization

The inclusion of dropout rates helps in preventing overfitting, making the model more robust.

Discussion

Limitations of the Project.

Computational Complexity

The process of hyperparameter tuning, especially with deep learning models, can be computationally expensive and time-consuming.

Overfitting Risk

Despite regularization techniques like dropout, there is still a risk of overfitting, especially if the model is too complex or the dataset is not sufficiently large.

Interpretability

Deep learning models, particularly those with multiple layers and complex architectures, can be difficult to interpret and understand.

Discussion

Potential Areas for Future Research.

Advanced Hyperparameter Tuning

Explore more advanced techniques for hyperparameter tuning, such as Bayesian optimization or genetic algorithms, to further improve model performance.

Model Interpretability

Investigate methods to improve the interpretability of deep learning models, such as attention mechanisms or model-agnostic interpretability techniques.

Transfer Learning

Explore the use of transfer learning to leverage pre-trained models and improve performance on specific tasks with limited data.

Data Augmentation

Implement data augmentation techniques to artificially increase the diversity of the training dataset and improve model robustness.

Ensemble Methods

Combine multiple models using ensemble methods to enhance overall performance and reduce the risk of overfitting.

Conclusion

Our Project Experience and Reflections.

This project guided the development and optimization of a deep learning model, focusing on hyperparameter tuning and performance evaluation to ensure its robustness and reliability.

Systematic Tuning

Using [RandomizedSearchCV] effectively explored a wide range of configurations, identifying the best model parameters.

Thorough Evaluation

Testing the model on diverse data, including edge cases, ensured its robustness in real-world scenarios.

Insightful Visualization

Plotting training and validation loss revealed key insights into the learning process, aiding in early detection of overfitting or underfitting.

References.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations (ICLR).