Recommendation System for Fashion Products

Introduction

With the increasing digitalization of the fashion retail industry, online stores and e-commerce platforms have become major channels for consumers to discover and purchase fashion products. However, the vast number of available products can overwhelm customers, leading to decision fatigue and abandoned shopping carts. Traditional methods of browsing, such as sorting by categories or popularity, do not fully leverage individual consumer preferences, resulted in reduction of customer satisfaction and engagement. This creates a need for a system that will enhance the overall shopping experience, reduce user churn, and increase sales on fashion e-commerce platforms.

Problem Identification

What opportunities exist to develop a personalized recommendation system that provide more accurate, diverse and personalized recommendations to users, thereby improving their shopping experience and increasing the likelihood of a purchase.

Data Wrangling/Data cleaning/Data munging

In this step, we performed a series of processes to explore, transform, and validate raw dataset retrieved into a high-quality and reliable data for analysis. This step include checking out following items:

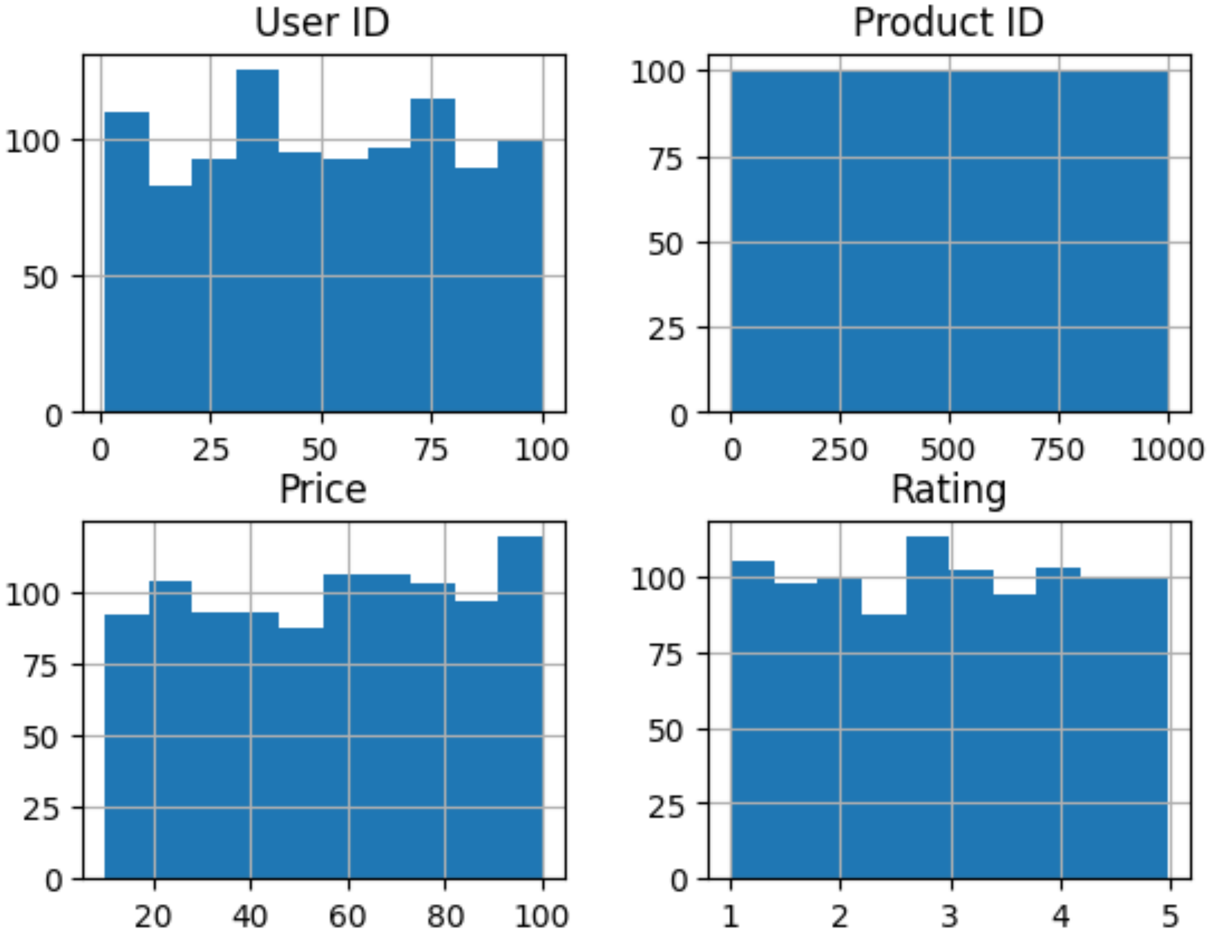
Missing values - No missing value has been detected

Outliers = No outliers detected in all numeric columns including User ID, Product ID, Price, Rating

Further visualized Numerical Data Distribution

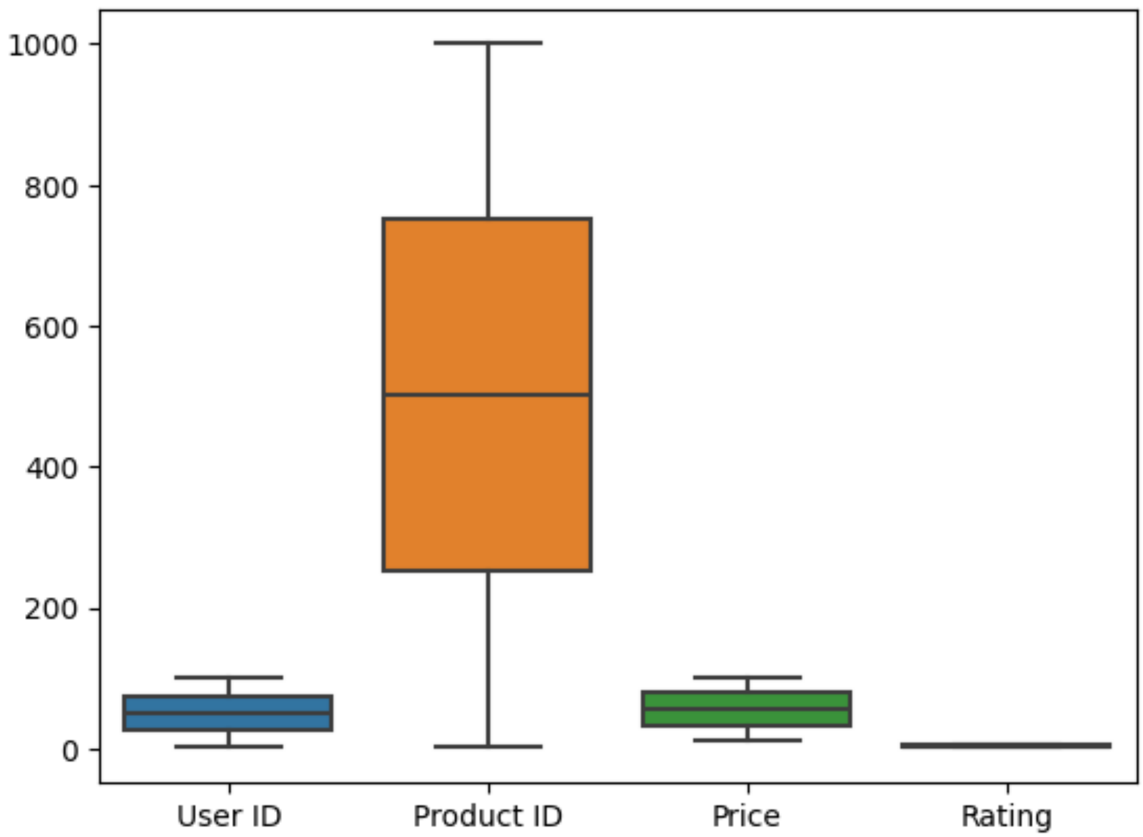
The distribution of data is multi-modal that has multiple peaks which means that data is not uniformly distributed and may represent **multiple distinct underlying processes** or

subgroups within the dataset.



Created a SeaBorn BoxPlot of the DataFrame to further confirmed if there are any outliers in any of the columns.

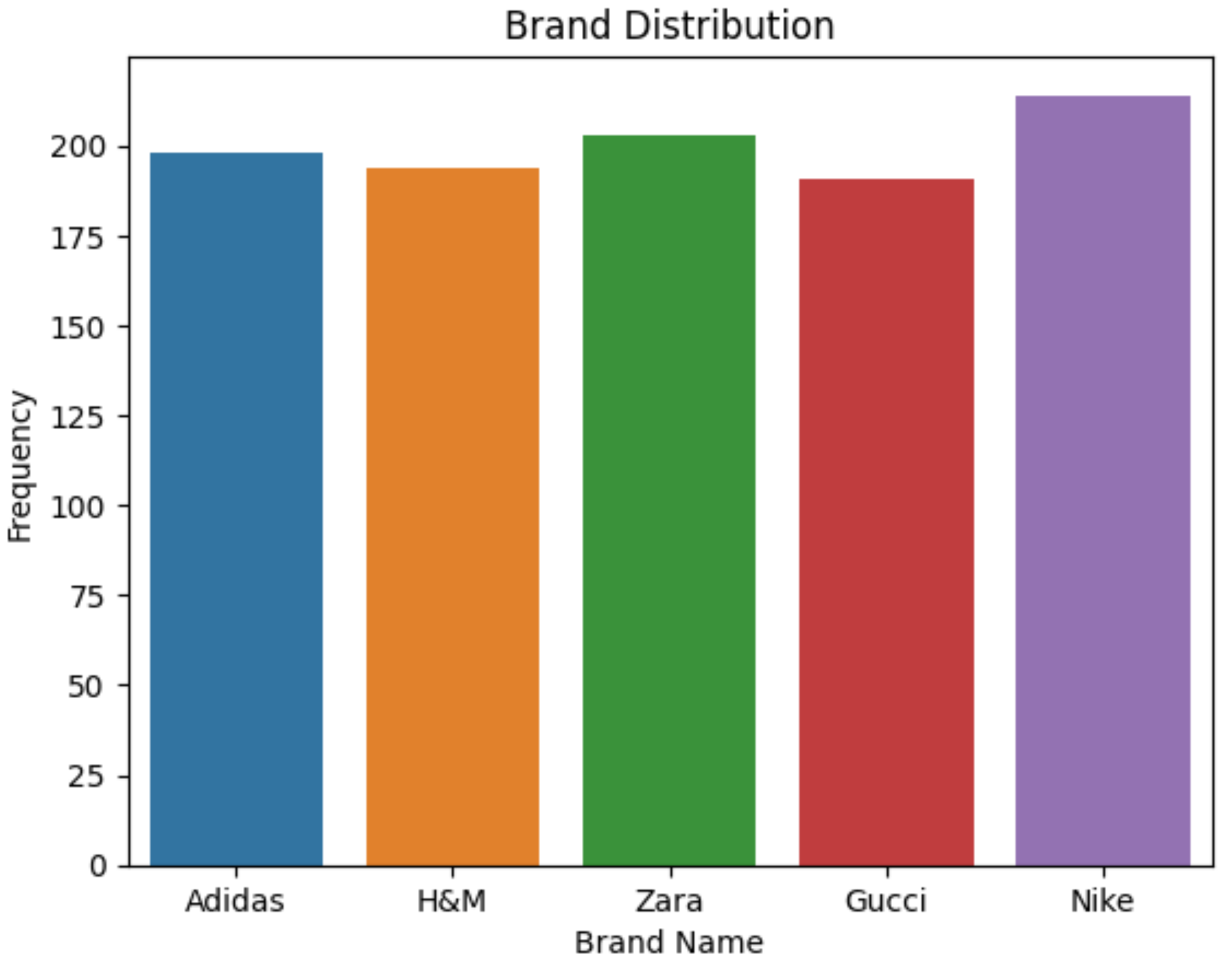
Result: No outliers detected in any of the numeric columns including User ID, Product ID, Price, Rating.

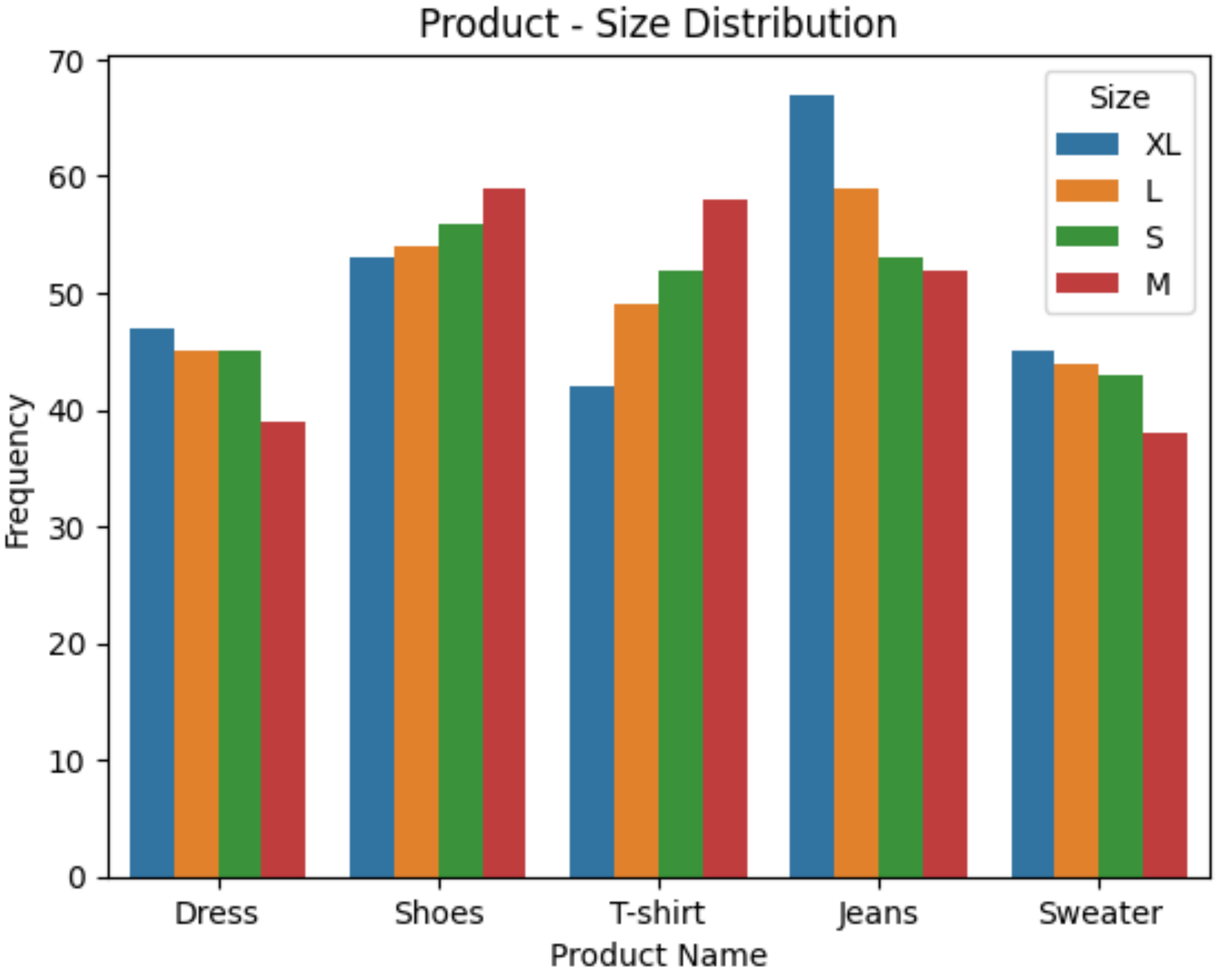


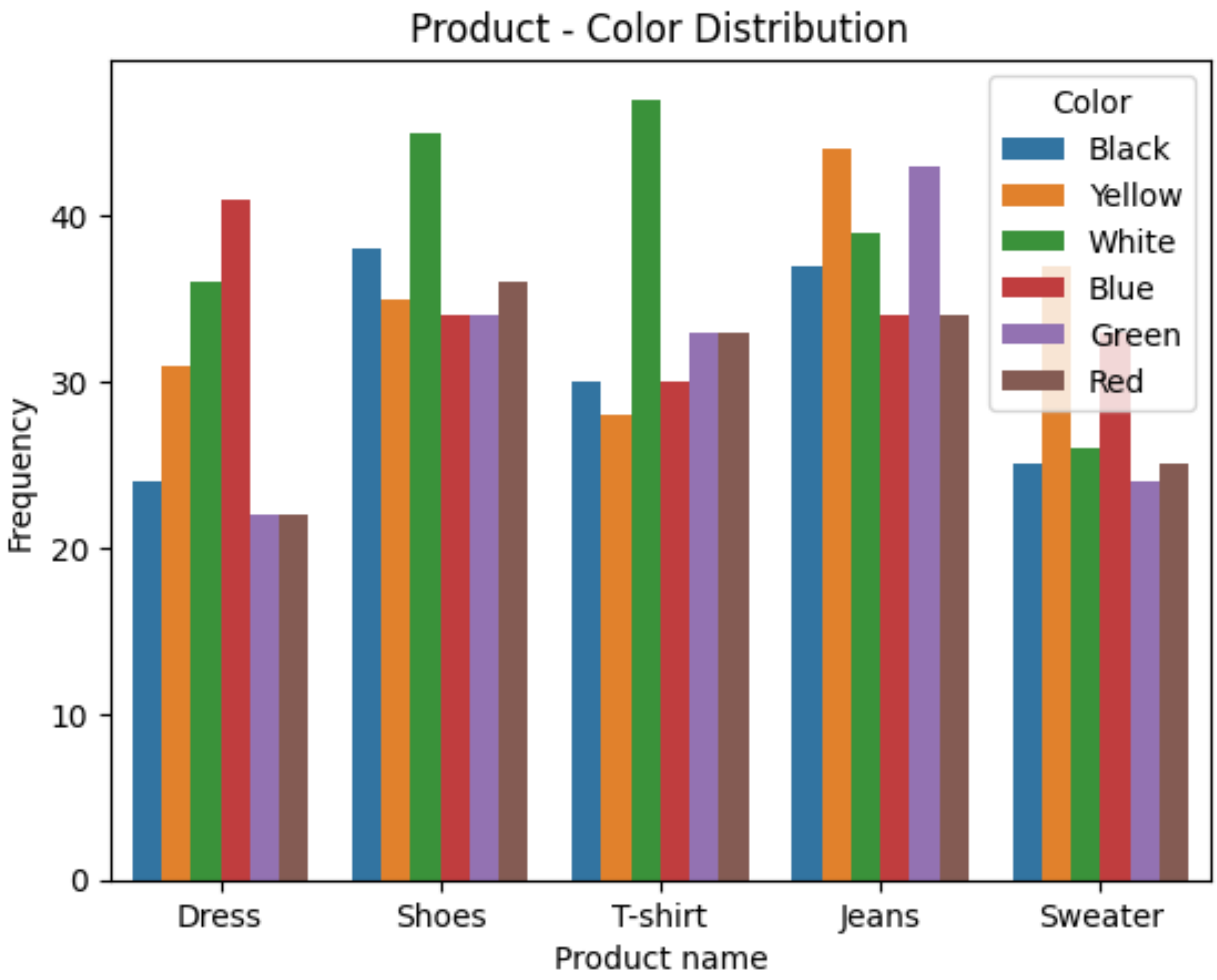
Further Visualized Categorical Data Distribution

Summary as follows:

- The frequency of Kid's fashion is slight higher than the men's and women's fashion.
- Among all the brands, the count of Nike products are slightly higher than other brands.
- Products such as Jeans, shoes and t-shirts are more popular in this dataset.
- In product jeans, XL seems most popular. M size tops the list in two product categories, shoes and t-shirts.
- Color White seems extremely popular in shoes and t-shirt categories.

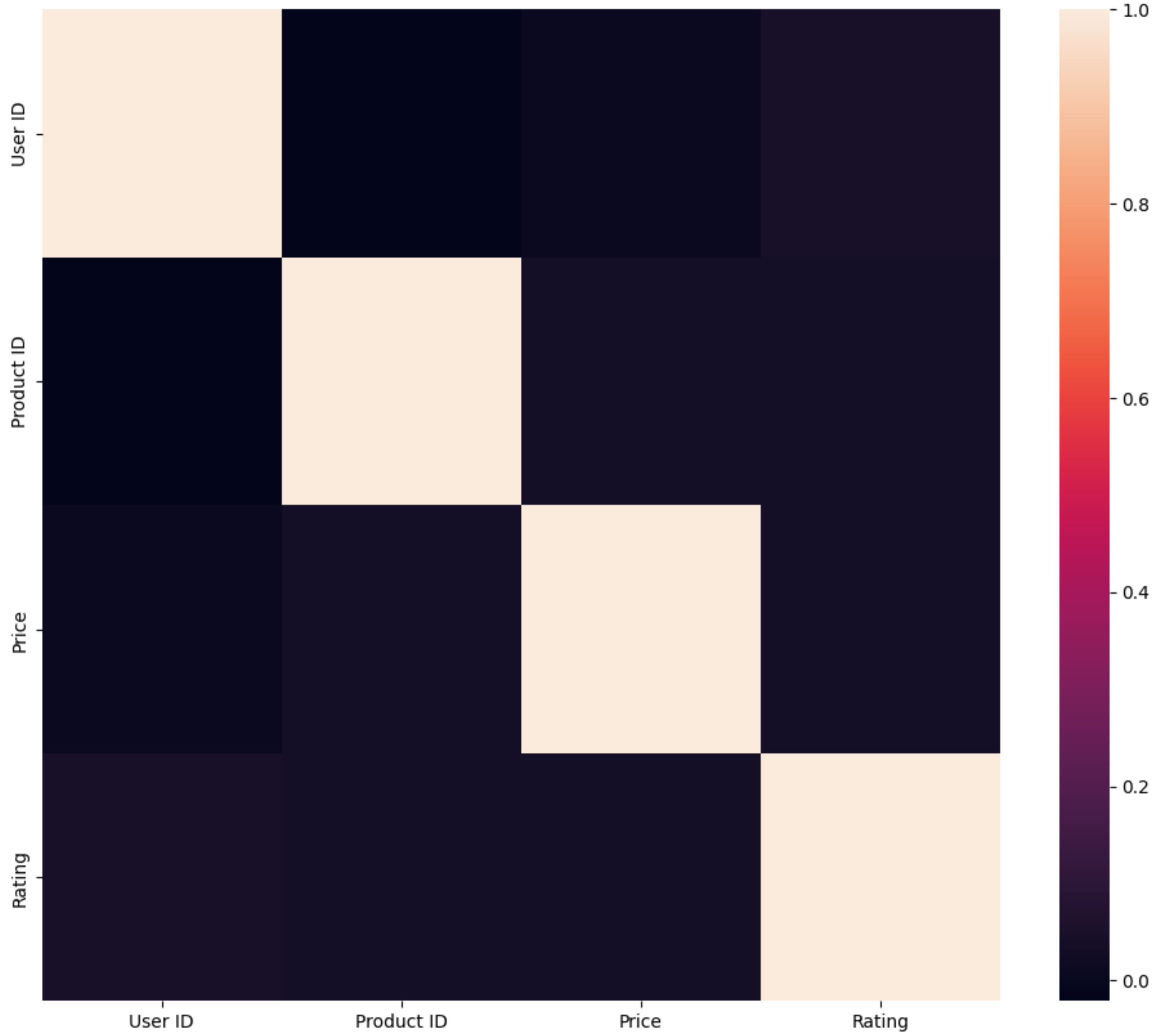

Brand Distribution
Frequency
200
175
150
125
100
75
50
25
0
Adidas
H&M
Zara
Gucci
Nike
Brand Name


Product - Size Distribution
Frequency
0
10
20
30
40
50
60
70
Dress
Shoes
T-shirt
Jeans
Sweater
Product Name
Size
XL
L
S
M


Product - Color Distribution
Color
Black
Yellow
White
Blue
Green
Red
Frequency
0
10
20
30
40
Dress
Shoes
T-shirt
Jeans
Sweater
Product name

Exploratory Data Analysis

Created **Feature correlation heatmap** to gain a high level view of relationships amongst the features and understand the underlying structure of the data. An interesting observations as there is no correlation between any of the features including the target variable “Rating”.



Statistical Analysis Summary

Tested multiple product attributes to see if they significantly affect product ratings using hypothesis testing and ANOVA.

Factor Compared	Test Type	p-value	Significant?	Conclusion
Men's vs Women's Fashion	t-test	0.520	❌ No	No significant difference in average ratings between categories
Price vs Rating	Pearson Correlation	0.284	❌ No	No significant correlation between price and rating
Brand vs Rating	ANOVA	0.156	❌ No	No brand stands out with significantly higher or lower ratings
Color vs Rating	ANOVA	0.727	❌ No	Color does not influence product rating
Size vs Rating	ANOVA	0.896	❌ No	No rating differences among product sizes

Insight: **User ratings are consistent across product features.** There is no statistical evidence that brand, price, size, color, or category significantly influence how users rate products in this dataset.

Pre-processing and Training data

To further improve our data quality and make the data useful for machine modeling.

Here is how our Target vs Feature Variables distributed:

- **Target Variables/ Label/ Dependent Variables** - Rating
- **Features/Independent Variables are as follows:** There are total 7 features.

Features	Feature Name
Numeric/Continous	Product ID - Identifier
Numeric/Continous	Price - Identifier

Features	Feature Name
Categorical	User ID
Categorical	Product Name
Categorical	Brand
Categorical	Category
Categorical	Color
Categorical	Size

Next step was to split the data into **training** and **test** set by partitioning the sizes with a 70/30 train/test split. This helps us learn the relationship between the input features and the target variable (Rating).

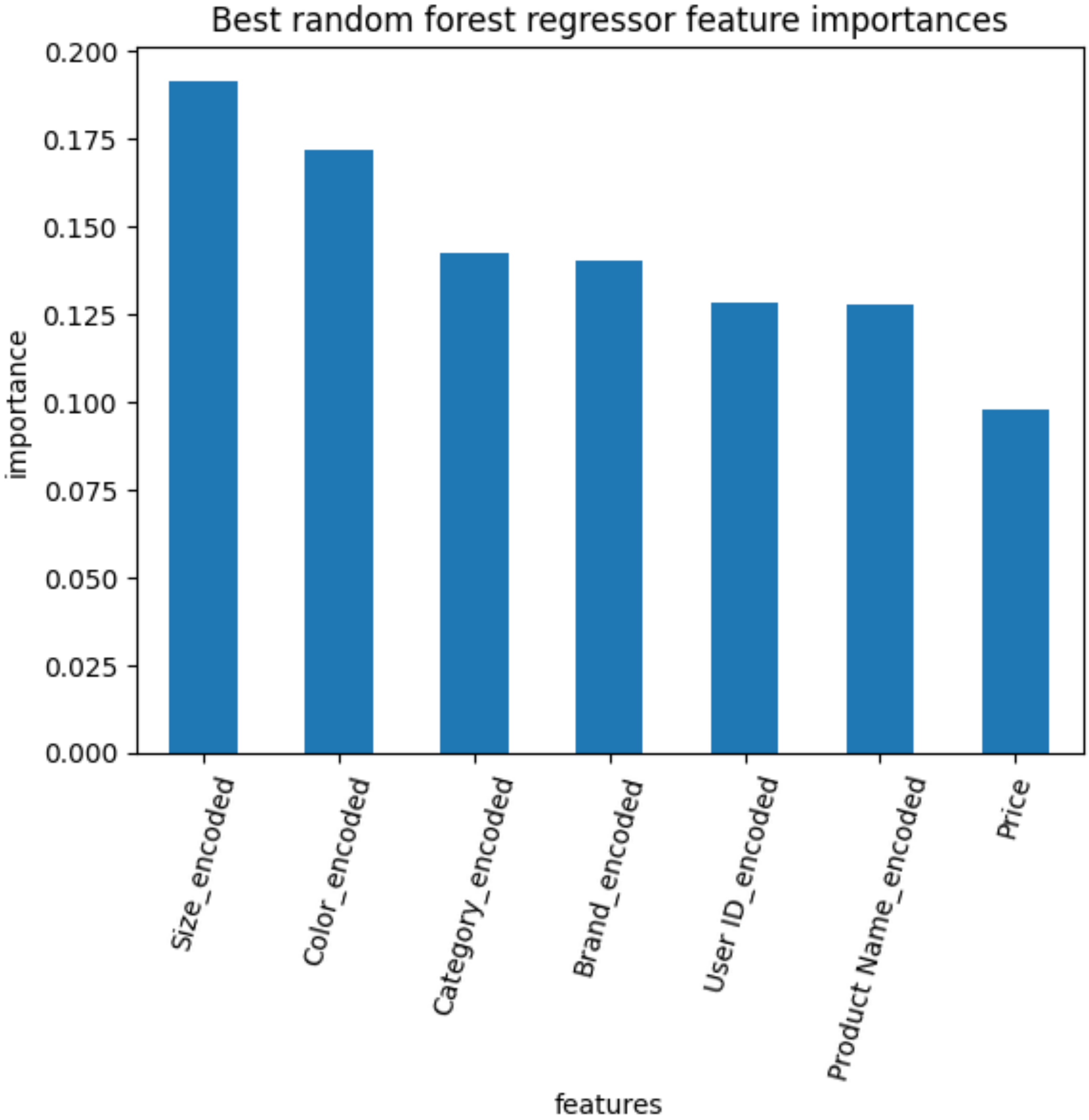
Modeling

In this step, I used four different methods:

1. Method 1 - Random Forest Model

Using Random forest regression model to identify the dominant top features, which are as follows:

Top Features	Importance
Size_encoded	0.191598
Color_encoded	0.171933
Category_encoded	0.142257
Brand_encoded	0.140161
User ID_encoded	0.128346
Product Name_encoded	0.127769
Price	0.097935

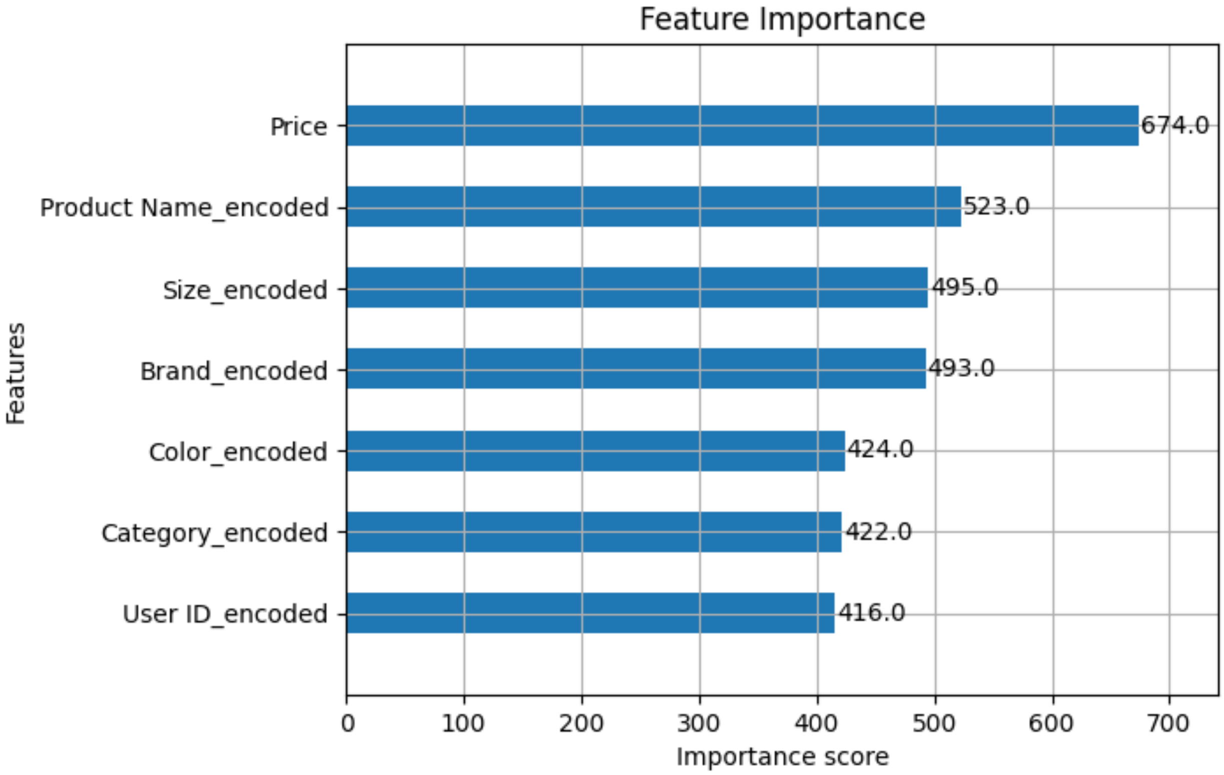


2. Method 2 - XG BOOST Model

XGBoost also provides a way to calculate, visualize and measure feature importance, including weight.

Using XG Boost model to identify the dominant top features, which are consistent with Random Forest model above are as follows:

Top Features	Importance (weight)
Price	674
Product Name_encoded	523
Size_encoded	495
Brand_encoded	493
Color_encoded	424
Category_encoded	422
User ID_encoded	416



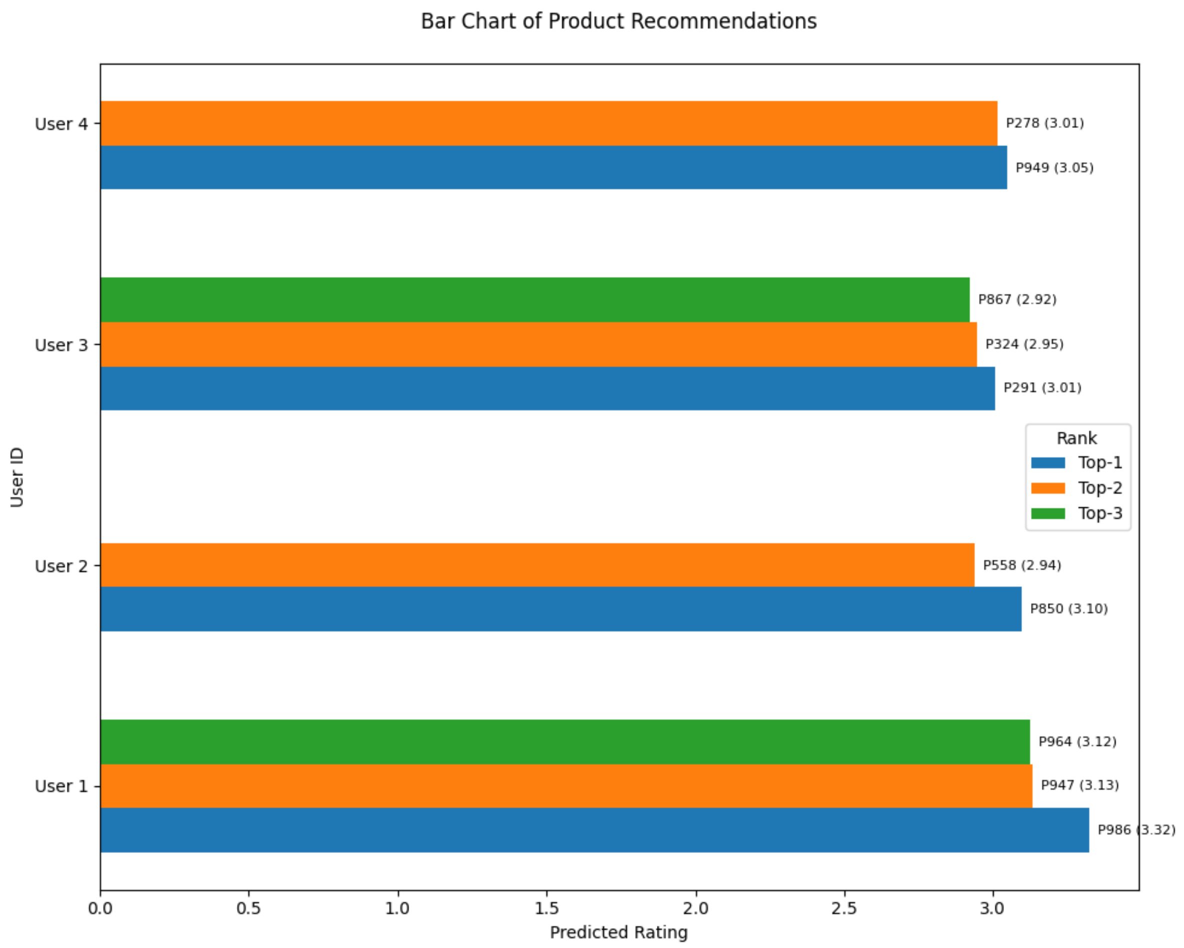
3. Method 3 - CATBOOST Model

Here we predict a target, ratings (a continuous variable) based on a variety of features (both numerical and categorical) features such as Price (numerical), Product Name, User ID, Category, Size, Color, Brand (categorical).

Selected and finalized the CatBoost model as best model because of the reasonable Root Mean Squared Error (RMSE) of 0.016 as compared to Random Forest and XGBoost model. Considering the Scale of the Target Variable (Rating), which is 1-5, RMSE of 0.016 seems quite reasonable.

Sort out the predicted ratings for each user in descending order and then Select the top N products with the highest predicted ratings.This gives us the Top-N recommended products for each user based on the predicted ratings.

User ID	Product ID	Predicted Rating
1	986	3.322825
1	947	3.132882
1	964	3.123750
2	850	3.097654
2	558	2.938357
3	291	3.008616
3	324	2.947347
3	867	2.922040
4	949	3.047553
4	278	3.014595



Using Cat Boost model to identify the dominant top features, which are consistent with Random Forest model above are as follows:

- Color
- Category
- Size
- Product Name
- Brand
- User ID
- Product ID
- Price

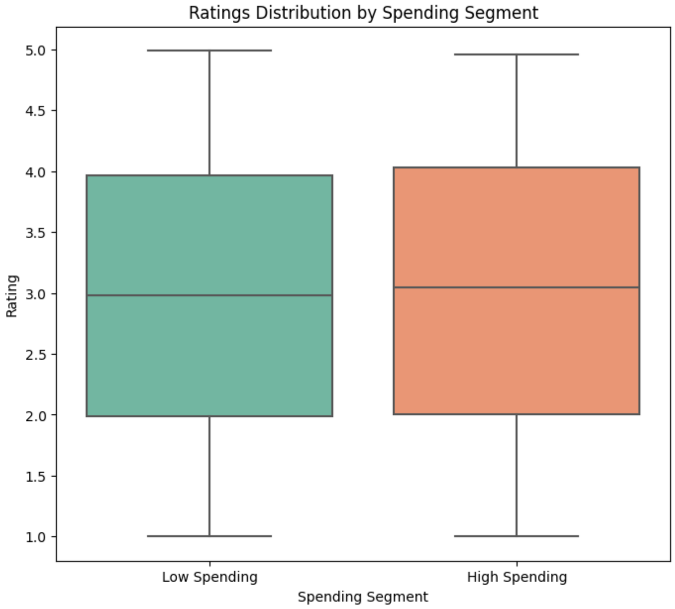
Visualize the Distribution of Ratings for Each Segment

This provide insights into how ratings vary across different segments in our case, high spenders vs low spenders.

Segmenting users into high-spending or low-spending group can further be used in data analysis and marketing strategy. By doing this, we can personalize marketing efforts, improve customer experience, and optimize business strategies.

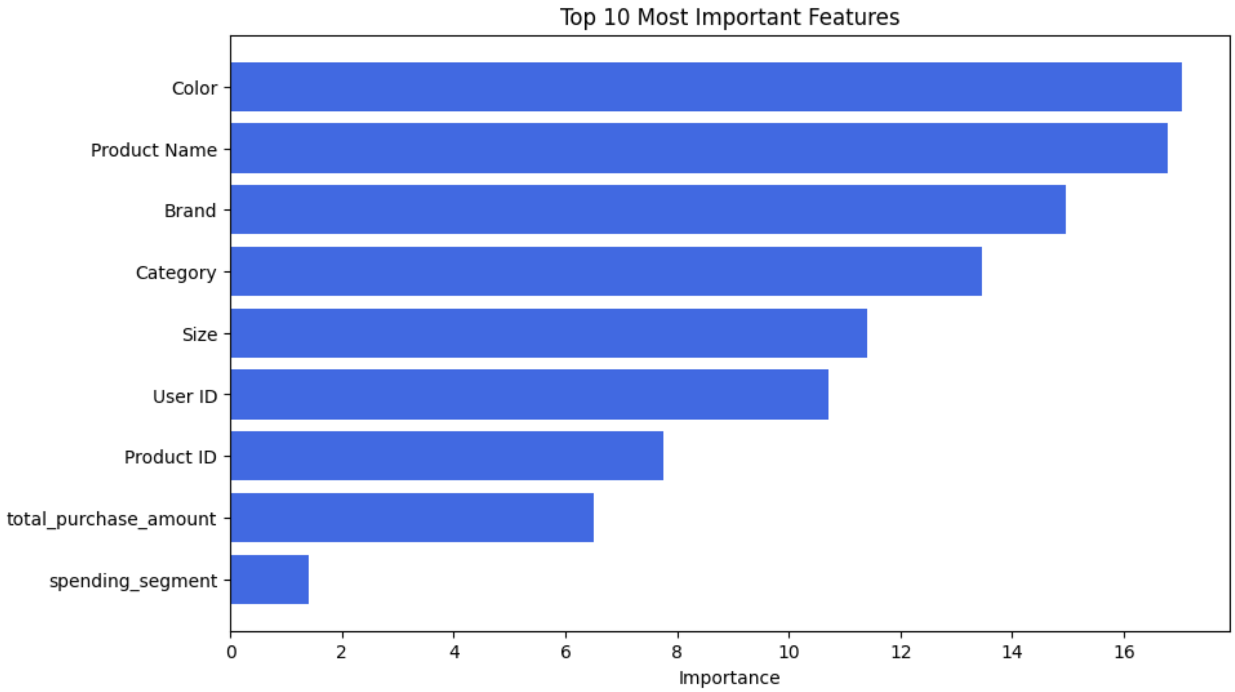
Here's I segment users based on these a key dimension — spending as how many users purchase the most products.

High-Spending: Users whose total spending or average purchase amount is above a threshold, 70.
Low-Spending: Users whose total spending or average purchase amount is below a certain threshold, which is 70.



Computed Average Rating per Spending Segment

Spending Segment	Rating
High - Spending	3.035180
Low - Spending	2.970891



Recommendations

This CatBoost model can be further improved by:

- Focus on top feature importances such as Product ID, Price, Size M and Color Yellow, which is consistent among all the models.

- Tuning the model periodically with fresh data to keep it up to date.

- Add more relevant features that may significantly affect the ratings by the user since no current attributes are affecting rating behavior.

- Can add features affects especially time-based features that track seasonal preferences (e.g., winter jackets in cold months).

- Recommend products based on recent purchases or trends.

- Can include external data sources such as marketing campaigns, competitor actions to stay updated.