

Capstone_Presentation

Recommendation for Fashion Product Project

Presented by - Swati Sharma

Context

Digitalization of Fashion Retail:

- Online stores and e-commerce platforms, major channels for discovering and purchasing fashion products.

Challenge of Overwhelming Product Choices:

- The vast number of available products can overwhelm customers.
- Leads to decision fatigue and abandoned shopping carts.

Limitations of Traditional Browsing Methods:

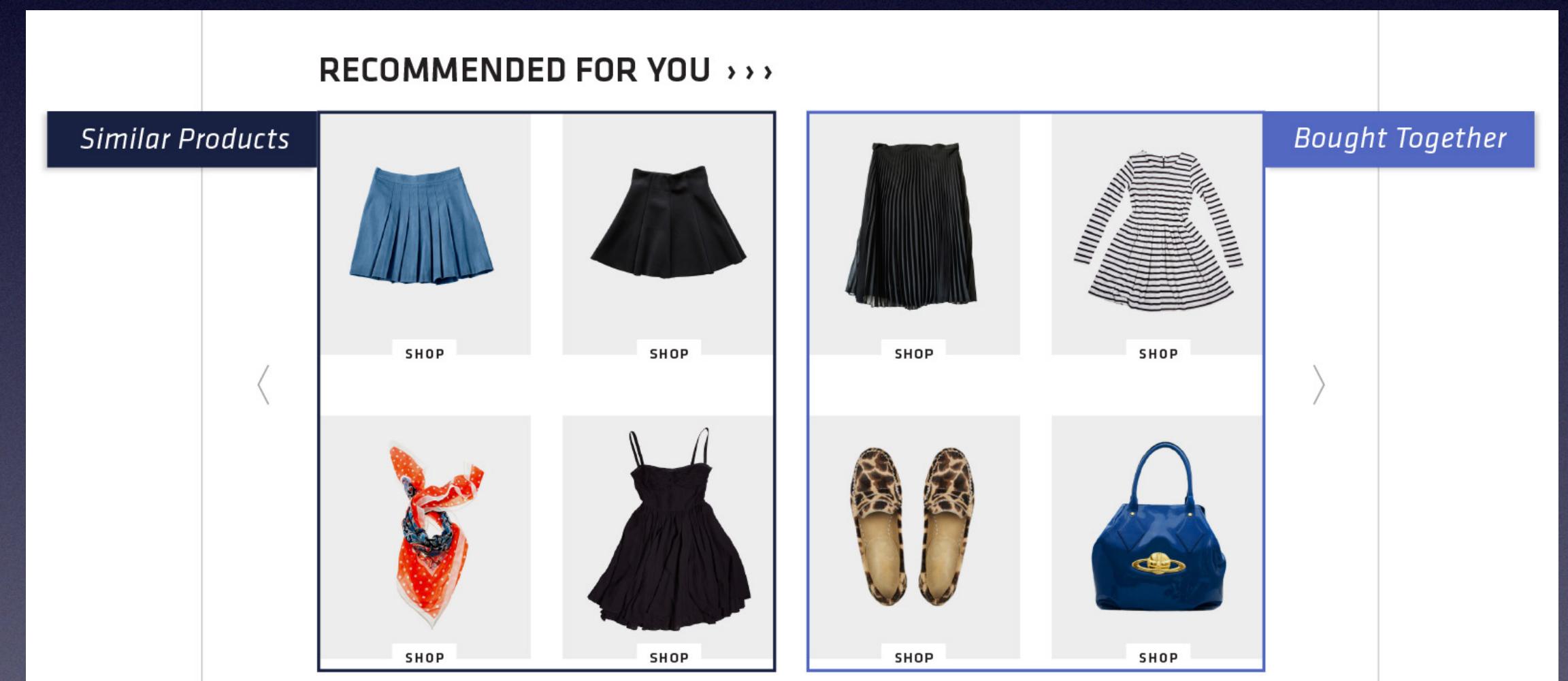
- Sorting by categories or popularity does not fully address individual consumer preferences.
- Results in reduced customer satisfaction and engagement.



What is the Problem?

What opportunities exist to develop a personalized recommendation system that not provide more accurate, diverse and personalized recommendations to users but also provide customer insights to the business.

- Results in enhancing the shopping experience, reduce user churn, and increase sales on fashion e-commerce platforms.
- Solve the problem of decision fatigue and improve the overall shopping experience.



Data Collection

Gather historical data: Collected raw data as a csv file with columns such as User ID, Product ID, Product name, Brand, Category, Price, Rating, Color, Size.

Data sources: Kaggle, an external data source.

	User ID	Product ID	Product Name	Brand	Category	Price	Rating	Color	Size
0	19	1	Dress	Adidas	Men's Fashion	40	1.043159	Black	XL
1	97	2	Shoes	H&M	Women's Fashion	82	4.026416	Black	L
2	25	3	Dress	Adidas	Women's Fashion	44	3.337938	Yellow	XL
3	57	4	Shoes	Zara	Men's Fashion	23	1.049523	White	S
4	79	5	T-shirt	Adidas	Men's Fashion	79	4.302773	Black	M
5	98	6	Dress	Adidas	Men's Fashion	47	1.379566	Yellow	L
6	16	7	Jeans	Gucci	Men's Fashion	37	1.356750	White	XL
7	63	8	Sweater	Zara	Kids' Fashion	64	4.360303	Blue	XL
8	96	9	Sweater	H&M	Men's Fashion	53	4.466182	Green	XL
9	36	10	T-shirt	Zara	Kids' Fashion	55	4.093234	White	XL

Goal: Create valuable insights that drive sales, engagement, and satisfaction
Structured data that include user behaviors, product attributes, and transaction details.

Independent vs Dependent Variables

User ID	Product ID	Product Name	Brand	Category	Price	Color	Size	Rating
19	1	Dress	Adidas	Men's Fashion	40	Black	XL	1.043159
97	2	Shoes	H&M	Women's Fashion	82	Black	L	4.026416
25	3	Dress	Adidas	Women's Fashion	44	Yellow	XL	3.337938
57	4	Shoes	Zara	Men's Fashion	23	White	S	1.049523
79	5	T-shirt	Adidas	Men's Fashion	79	Black	M	4.302773

Identifier

Independent
Variables

Dependent
Variable

Features/
Independent
Variables

Variables that
is changed

Target Variables/
Label/ Dependent
Variables

Variables
affected
by the change

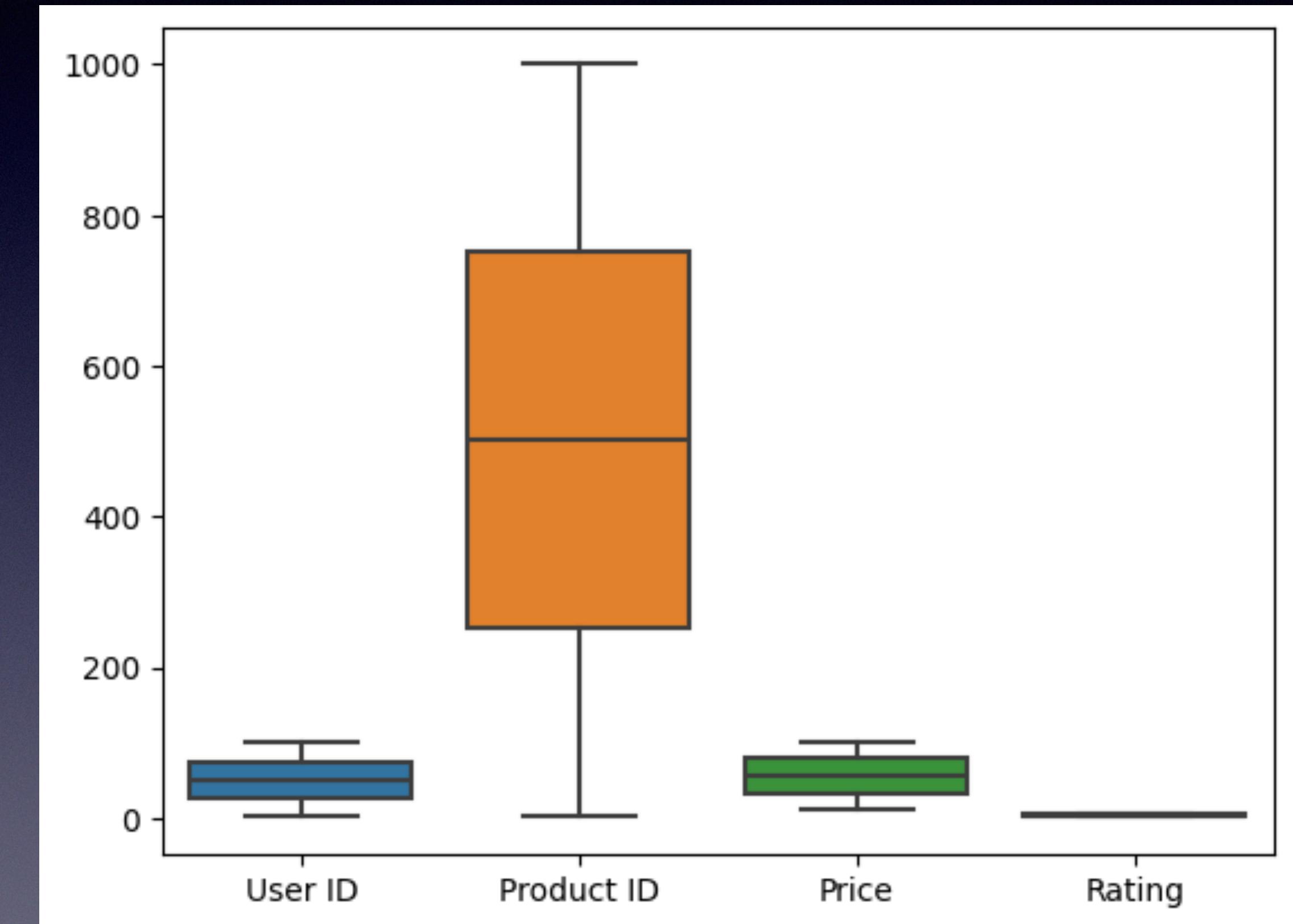
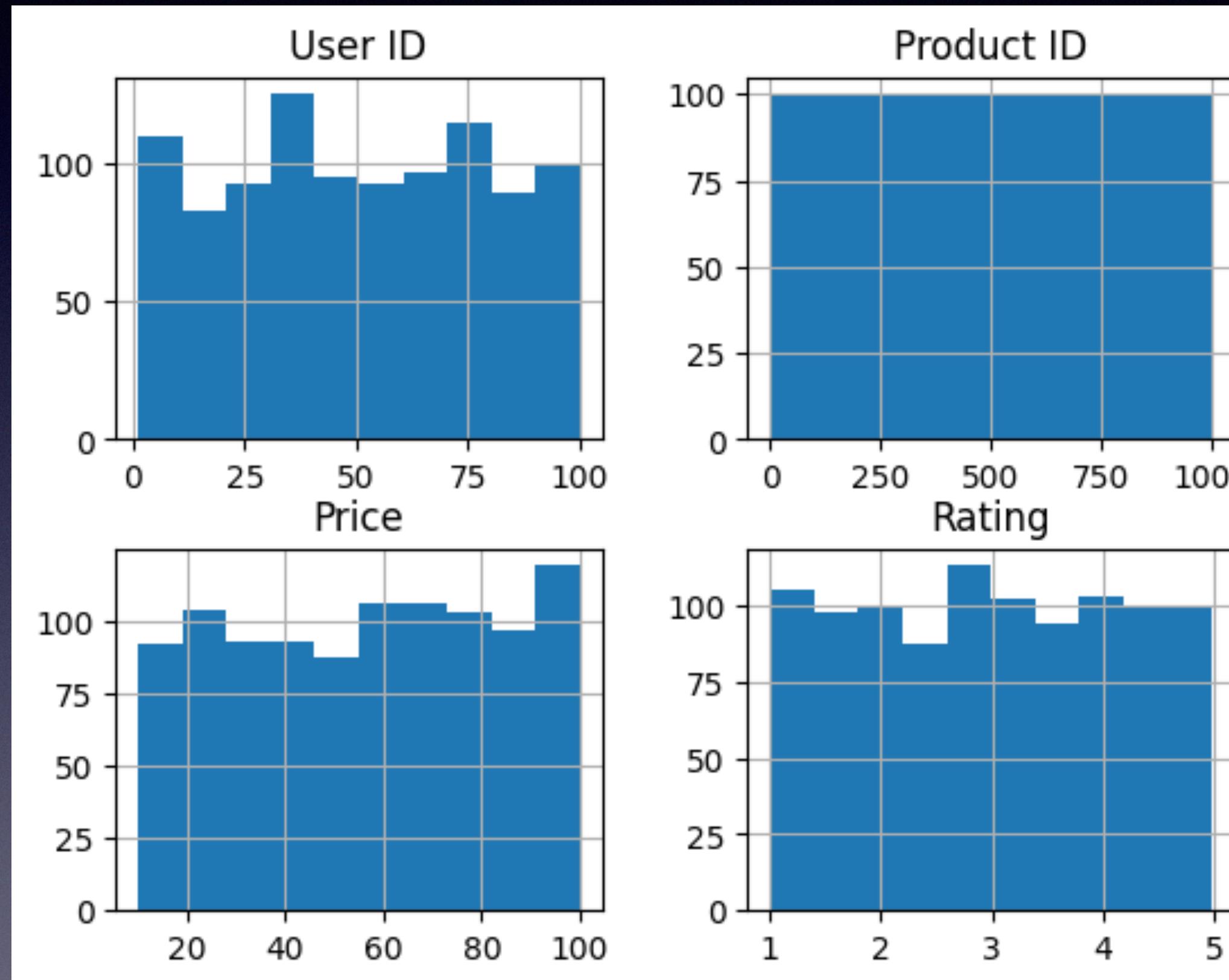
Data Wrangling / Data Cleaning

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



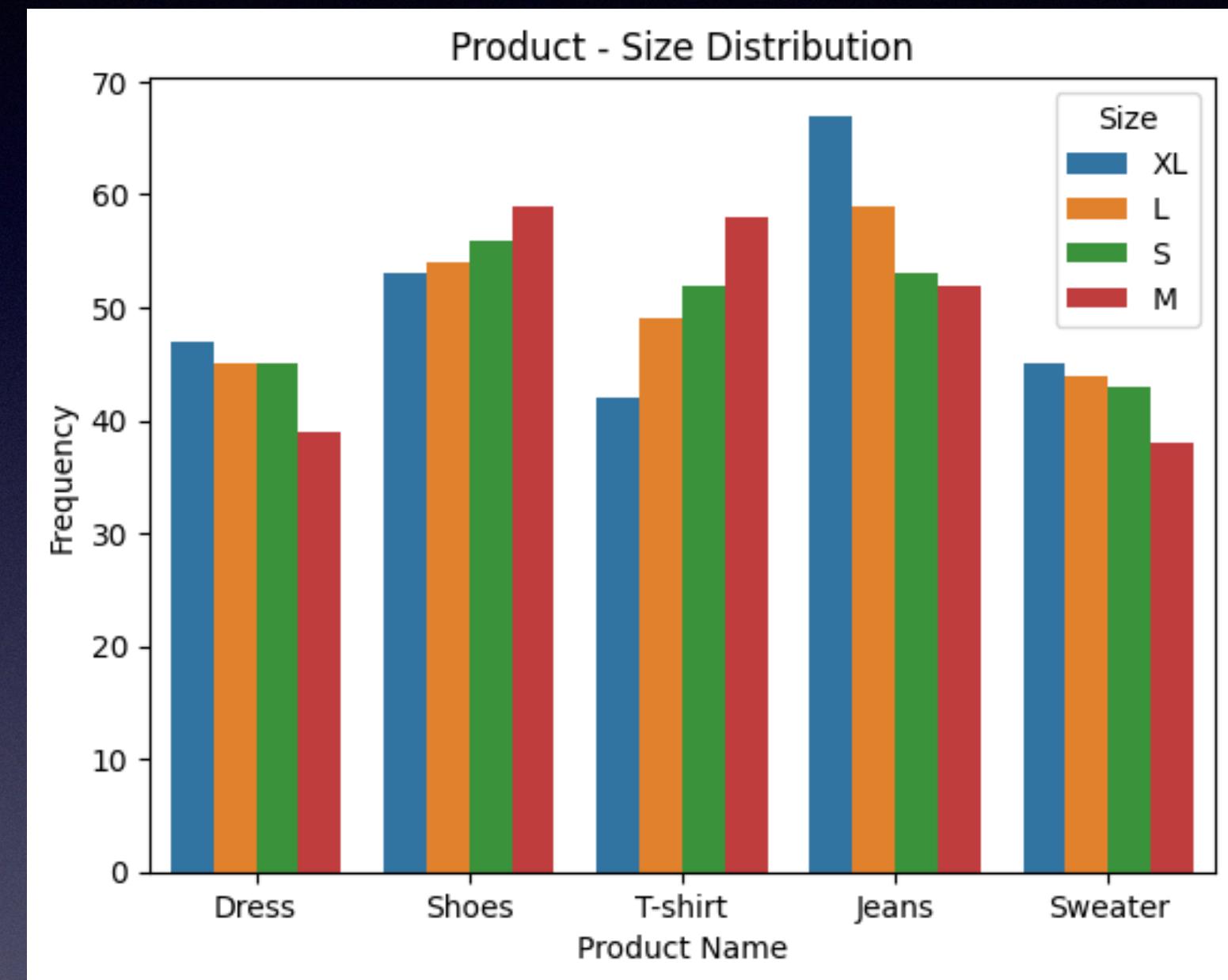
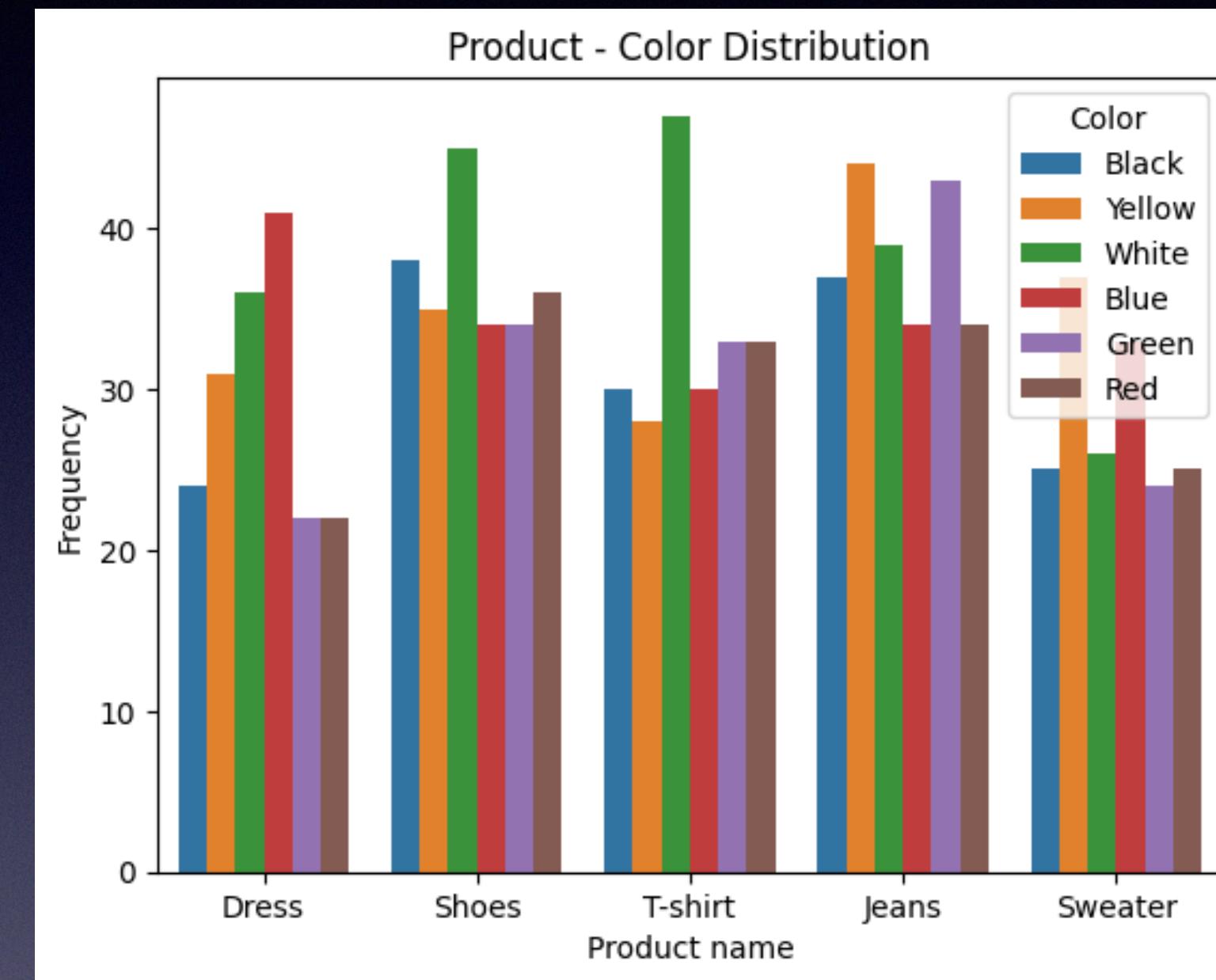
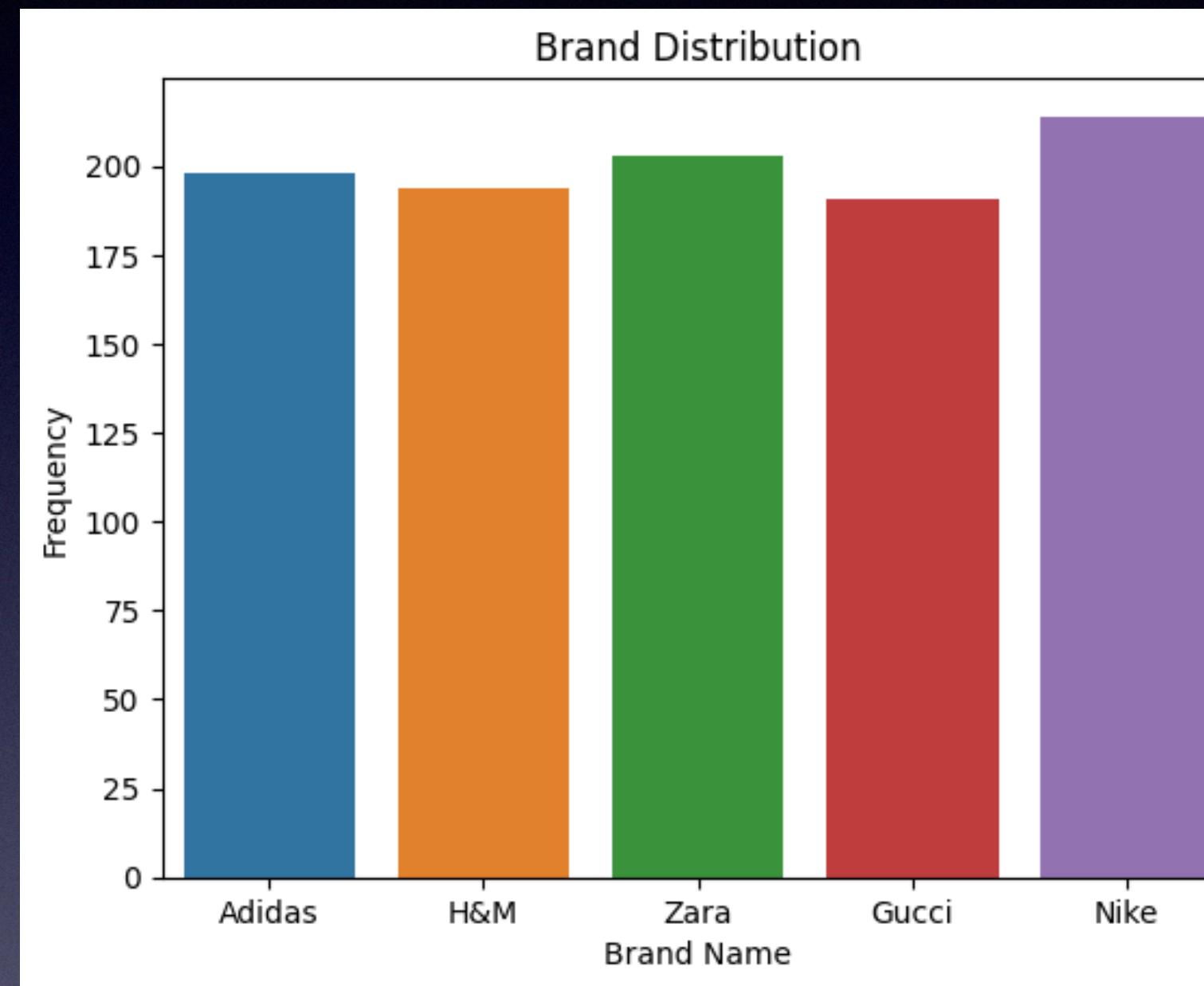
Visualize Numerical Data Distribution



Data Distribution is **multi-modal** with multiple peaks.

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

Visualize Categorical Data Distribution



Summary:

Frequency of Kid's fashion is slightly higher than the men's and women's fashion.

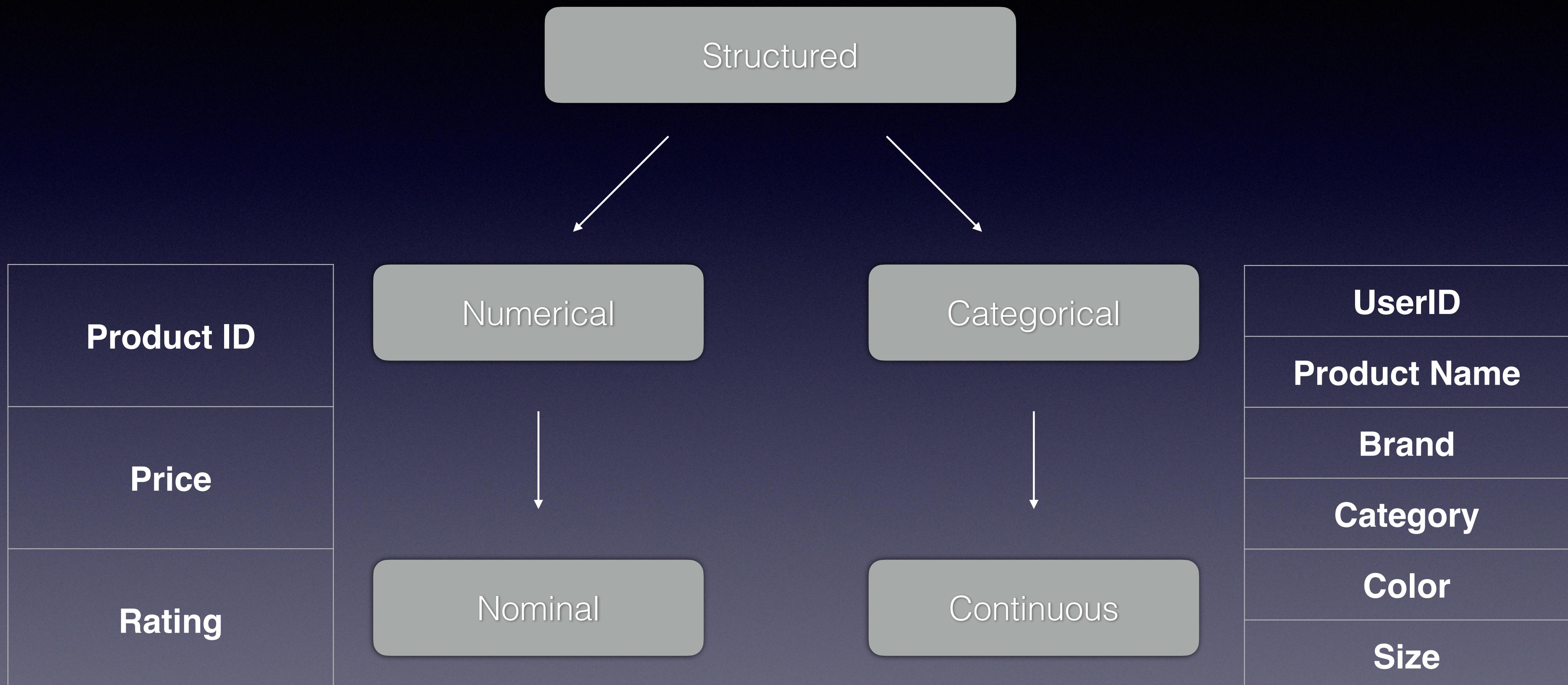
Count of Nike products are slightly higher than any other brands.

Products such as Jeans, shoes and t-shirts are more popular in this dataset.

In product type 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.

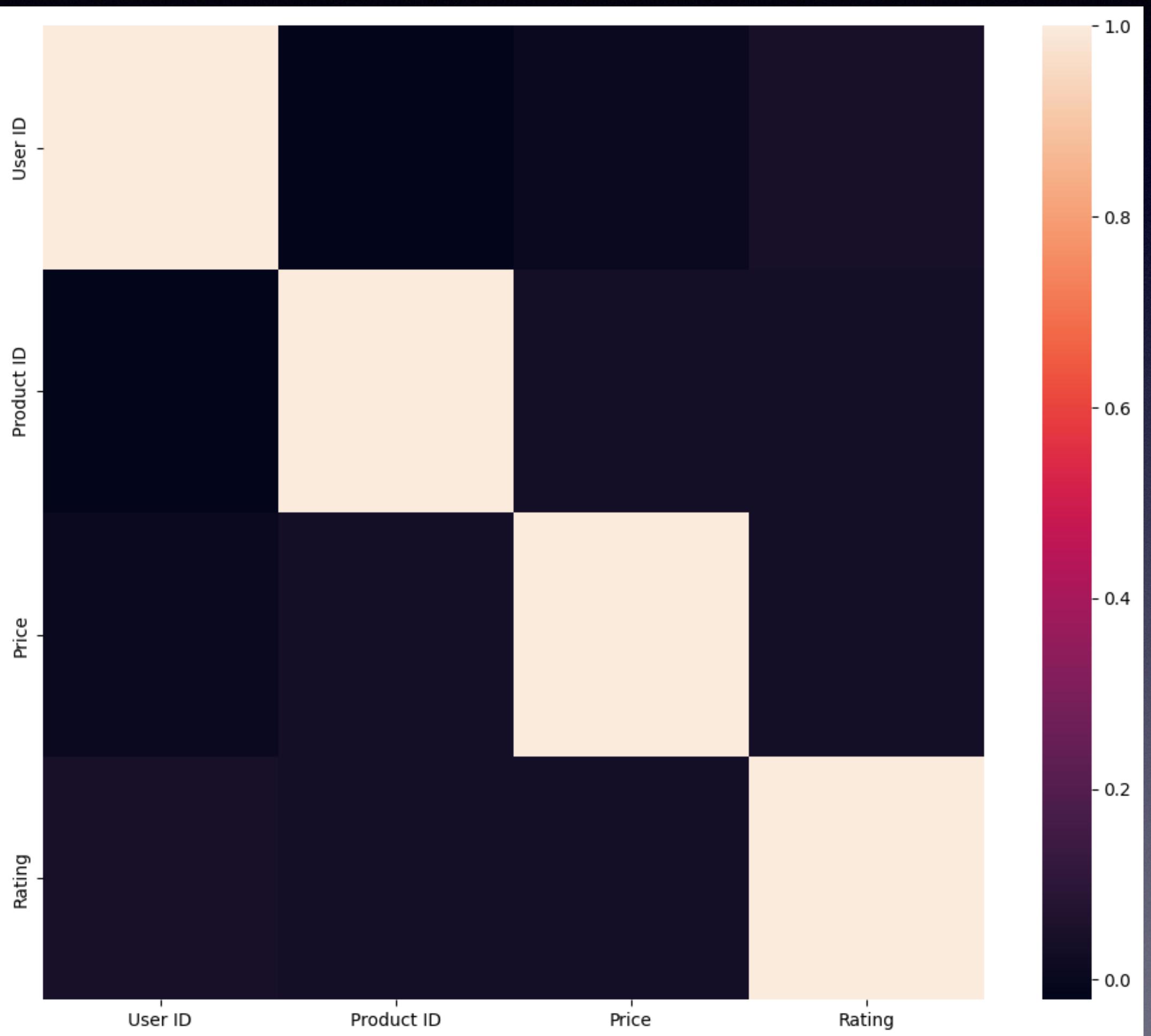
Type of Dataset



Exploratory Data Analysis

Feature correlation heatmap - Gain a high level view of relationships amongst the features

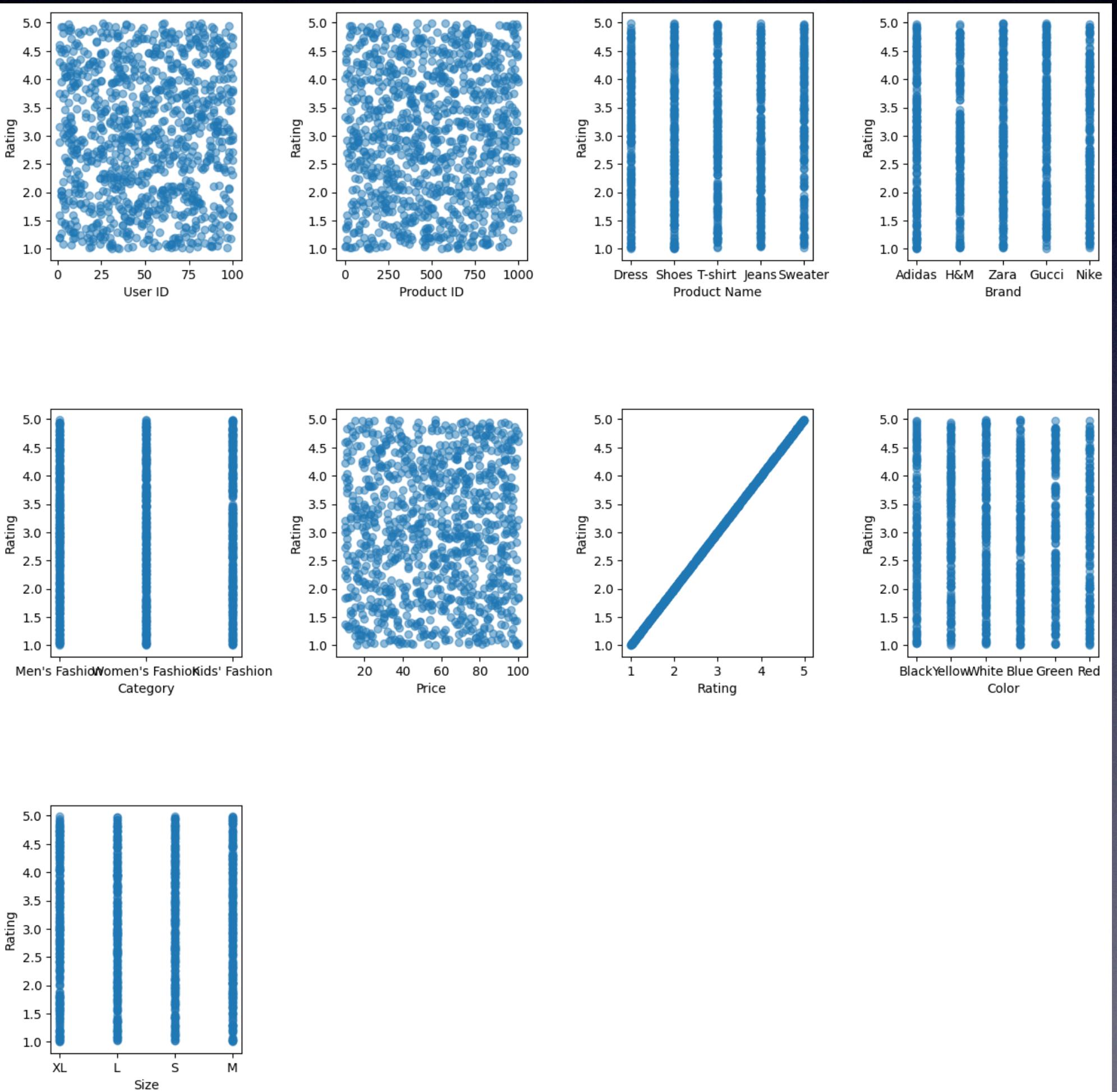
- There is no correlation between any of the features including the target variable “Rating”.
- Suggests non-linear relationships between the features.



Exploratory Data Analysis - Continue

Created a Scatter plots for visualizing the relationship between a numeric feature against target variable, Rating.

- Further confirms no clear relationship between various features.
- User ID, Product ID and Price seems too concentrated and no meaningful insight can be extracted.
-



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.

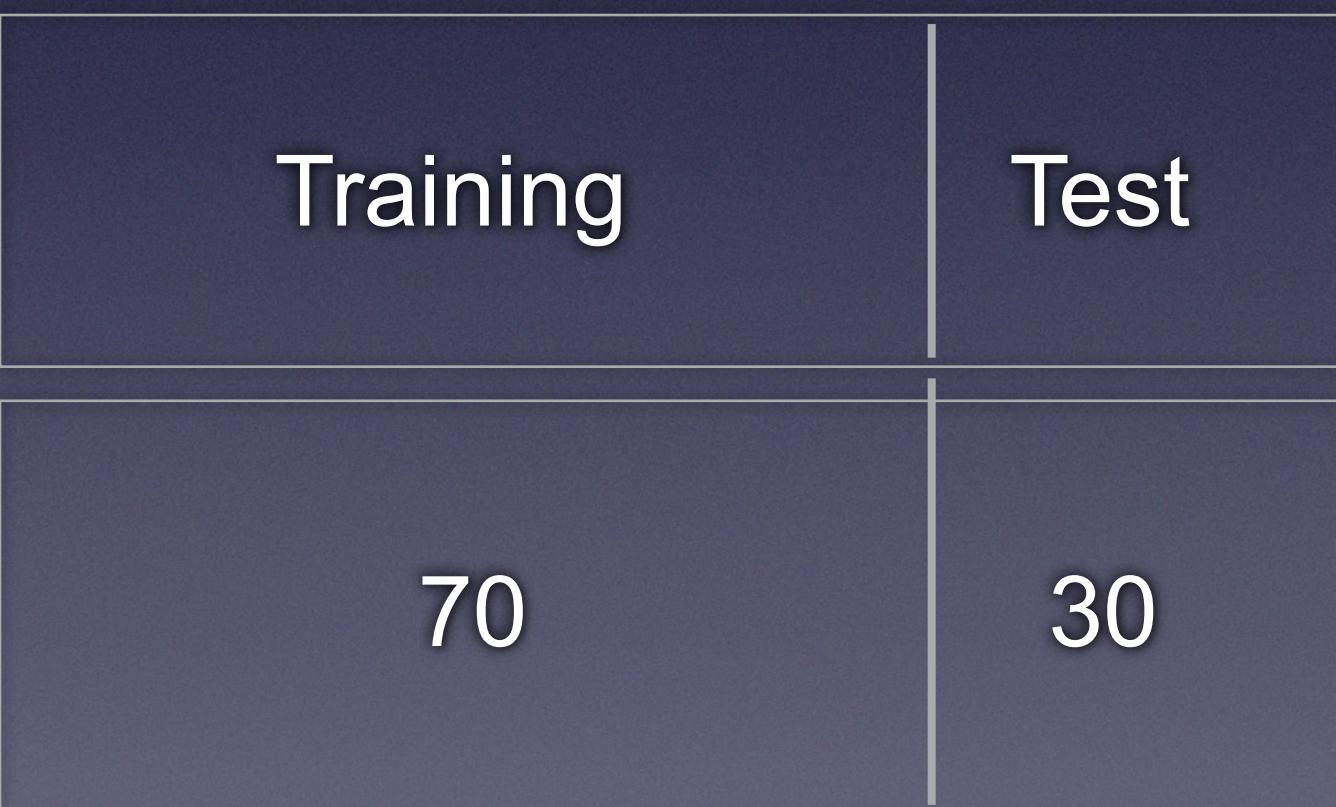
Pre-processing and Training data

Help us learn the relationship between the input features and the target variable (Rating).

Use CatBoostEncoder, target-based categorical encoder to encode Categorical Features:

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

Split Dataset



Machine Learning Algorithms

Choosing the right ML algorithm is crucial.

The choice depends on Problem Type:

- **Type of problem: Regression Problem**
(Target, Rating is continuous)
- **Nature of the data: Supervised learning as Labeled data**
- **Goals of the analysis: Predict a continuous numerical value**
- **Model performance: Mean Squared Error (MSE), R-squared (R²) and RMSE**

Target variable
(continuous)

Regression

Non -Linear
Relationship

Random Forest
Regressor

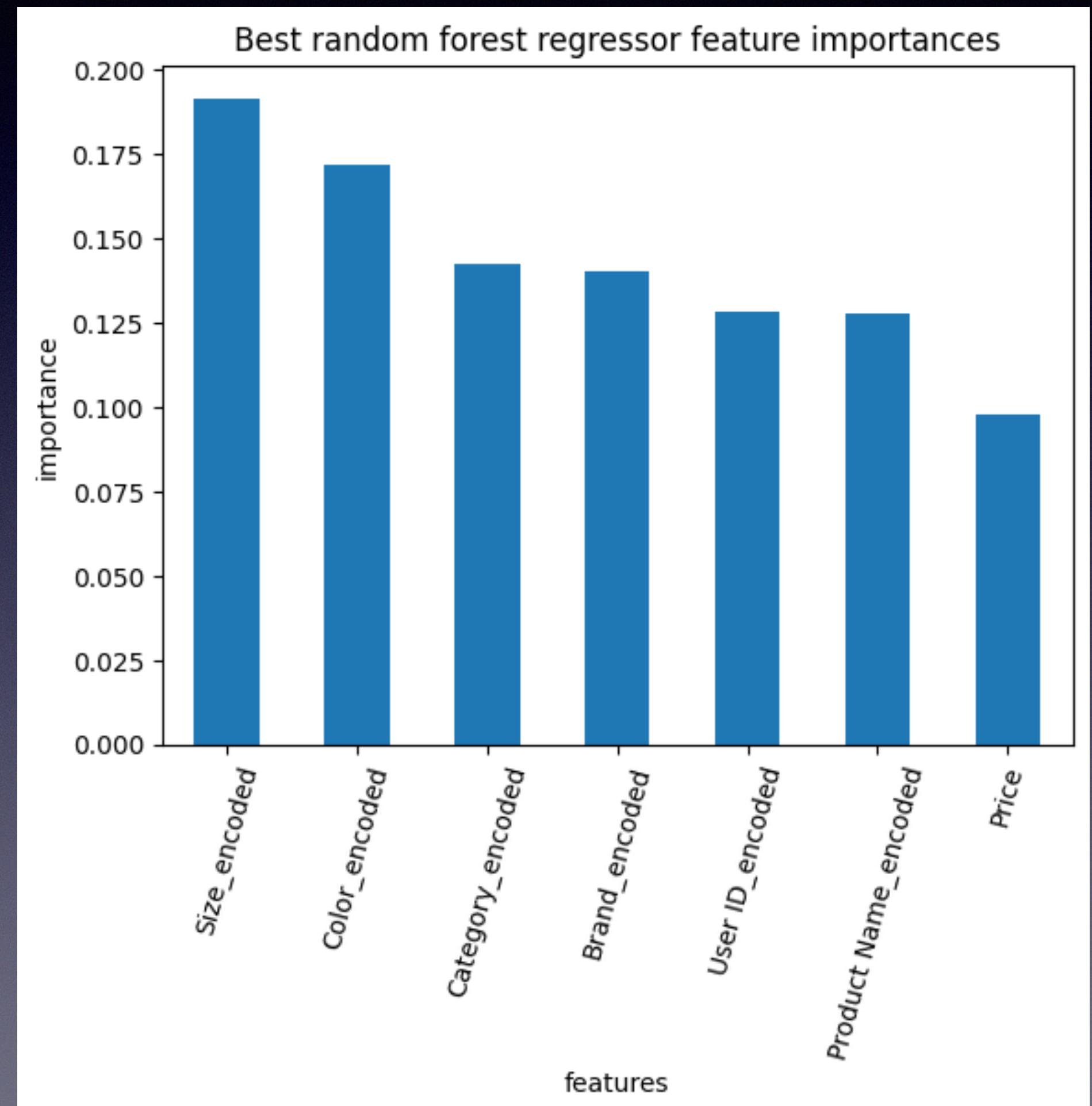
XGBoost

CatBoost

Random Forest Model

Identify the dominant features based on their importance in a Random Forest model are as follows:

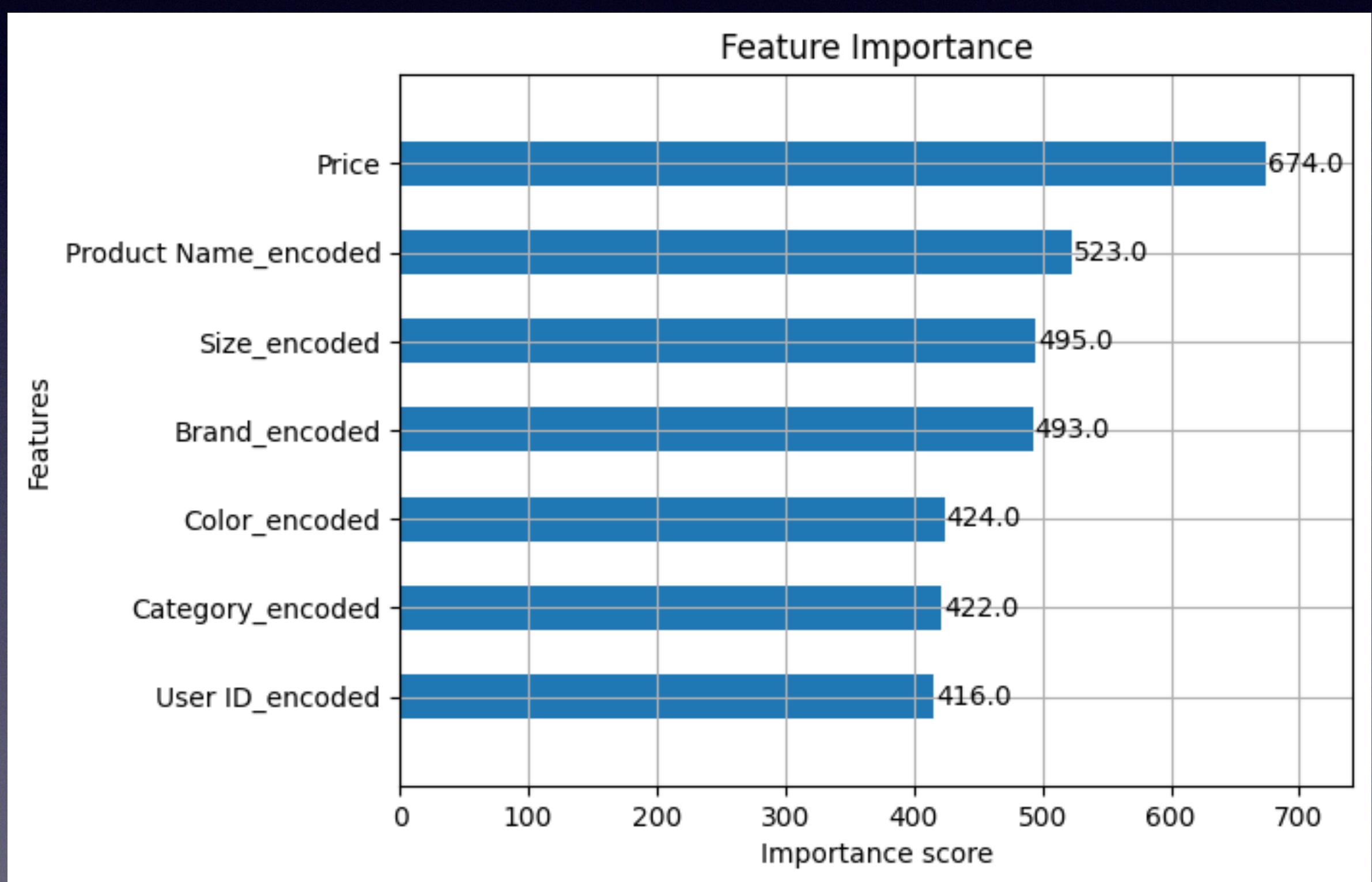
- Size
- Color
- Category
- Brand
- User ID
- Product Name
- Price



XG Boost Model

Identify the dominant features based on their importance in a XG Boost model are as follows:

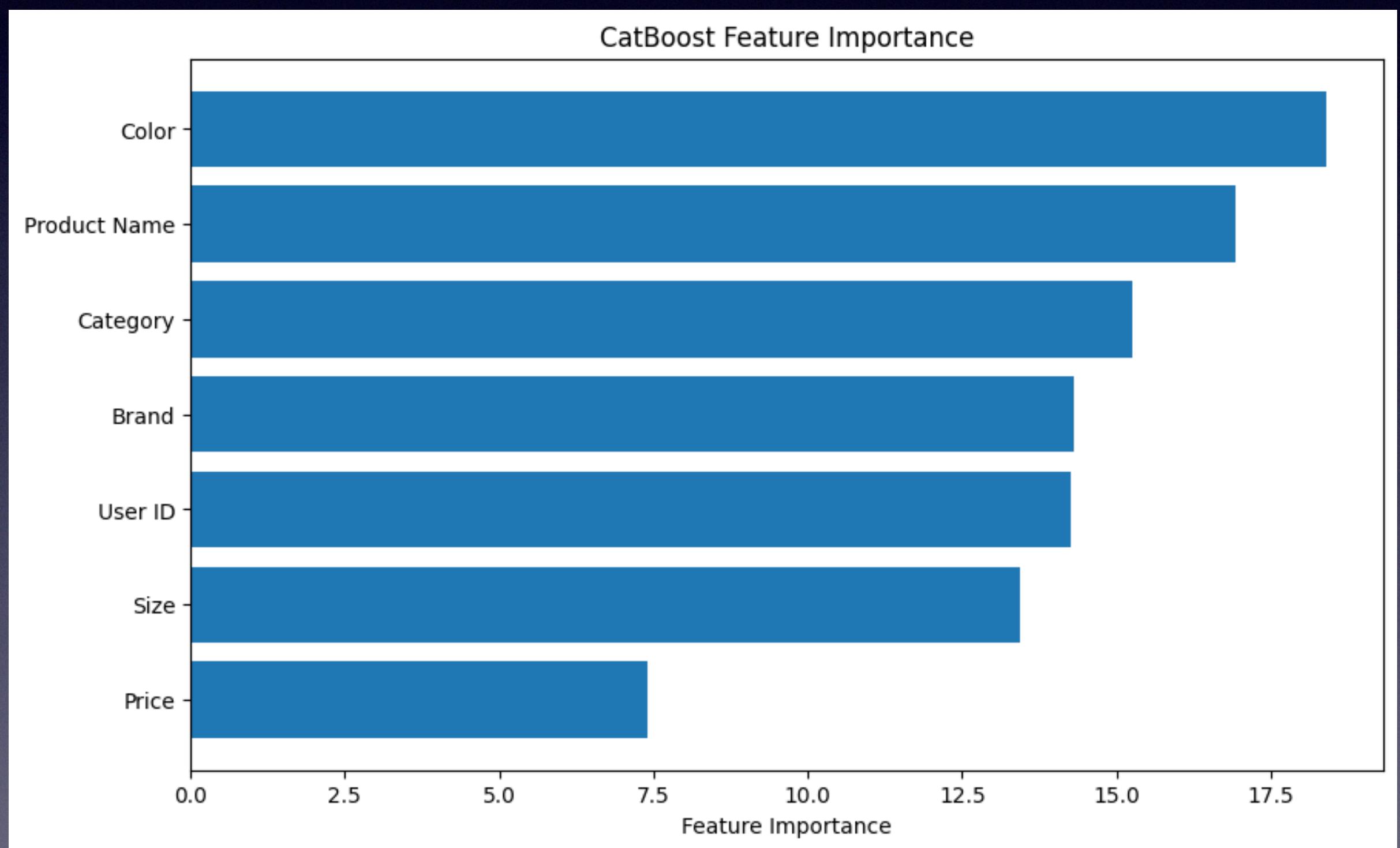
- Price
- Product Name
- Size
- Brand
- Category
- User ID



CatBoost Model

Identify the top dominant features based on their importance in a Cat Boost model are as follows:

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



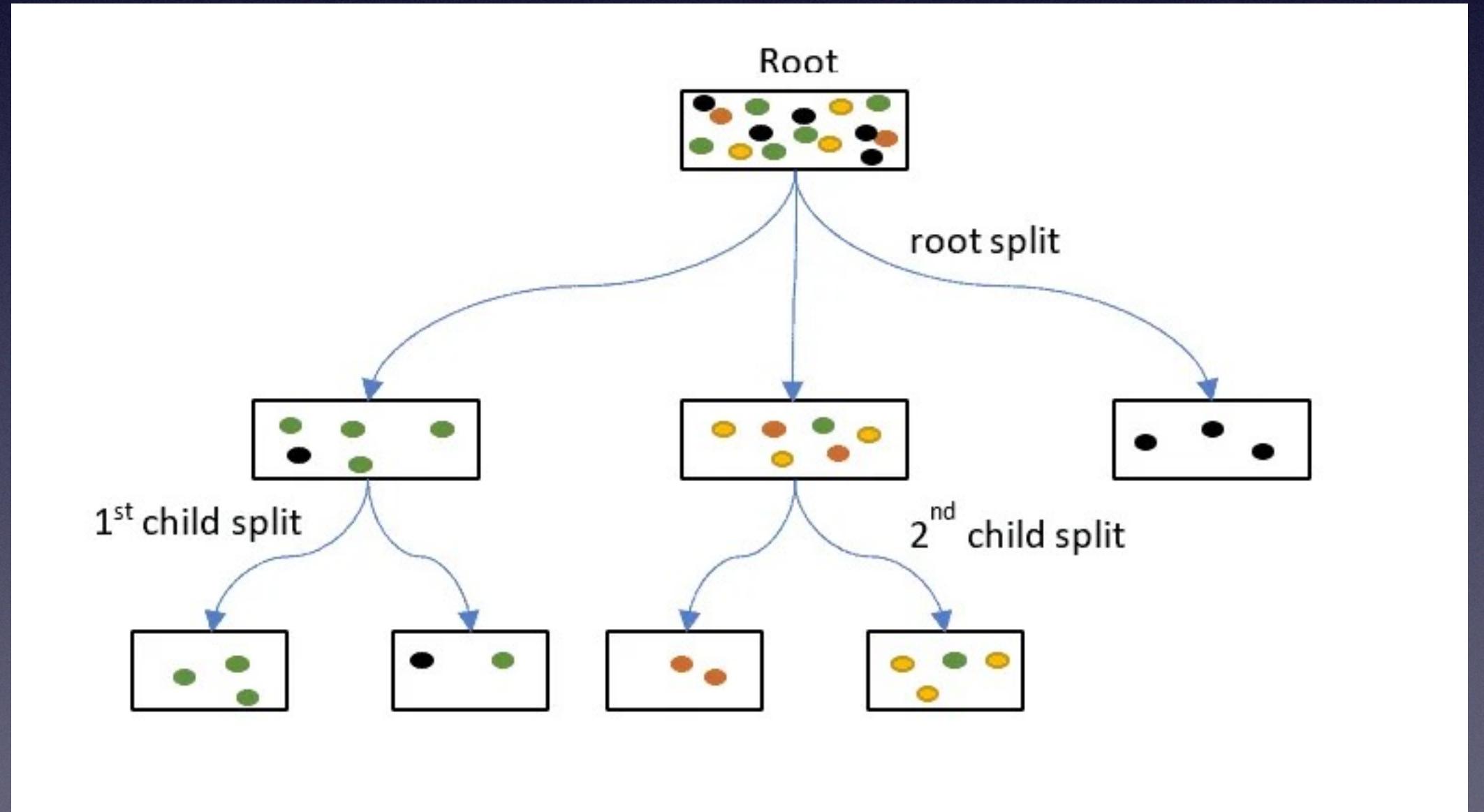
ML steps

	Random Forest	XGBoost	Cat Boost
Import libraries	scikit-learn	xgboost	scikit-learn
Initialize the model	RandomForestRegressor()	xgb.XGBRegressor()	CatBoostRegressor()
Train the model	model.fit(X_train, y_train)	model.fit(X_train, y_train)	model.fit(X_train, y_train)
Predict on test data	model.predict(X_test)	model.predict(X_test)	model.predict(X_test)
Evaluate the model	Mean Squared error: 1.26 Mean Absolute Error: 0.96 RMSE: 1.12	Mean Squared error: 1.50 RMSE: 1.22	RMSE: 0.02 (Post-Hyperparameter Tuning)
Hyperparameter Tuning	GridSearchCV	GridSearchCV	GridSearchCV

CatBoost Algorithm Overview

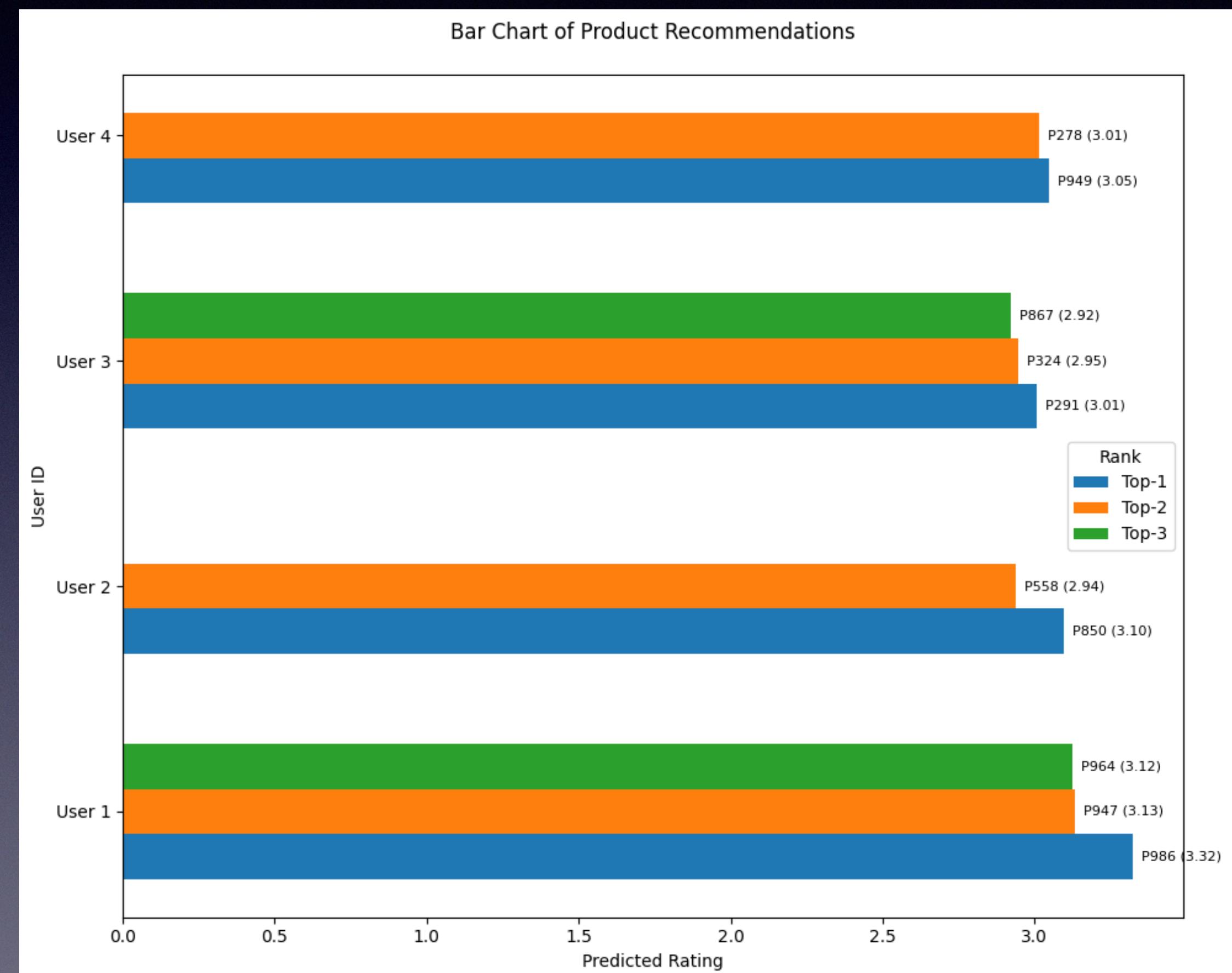
Key Features:

- An advanced **gradient boosting** algorithm optimized for performance, particularly when handling **categorical data**.
- Dataset include large number of categorical features such as Color, Size, User ID, Category etc.
- Making it an excellent choice for our training dataset
- Its ability to handle categorical variables directly and its **ordered boosting** method provide a significant advantage over other algorithms like **XGBoost** and **Random Forest Model**.



Top-N recommended products

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

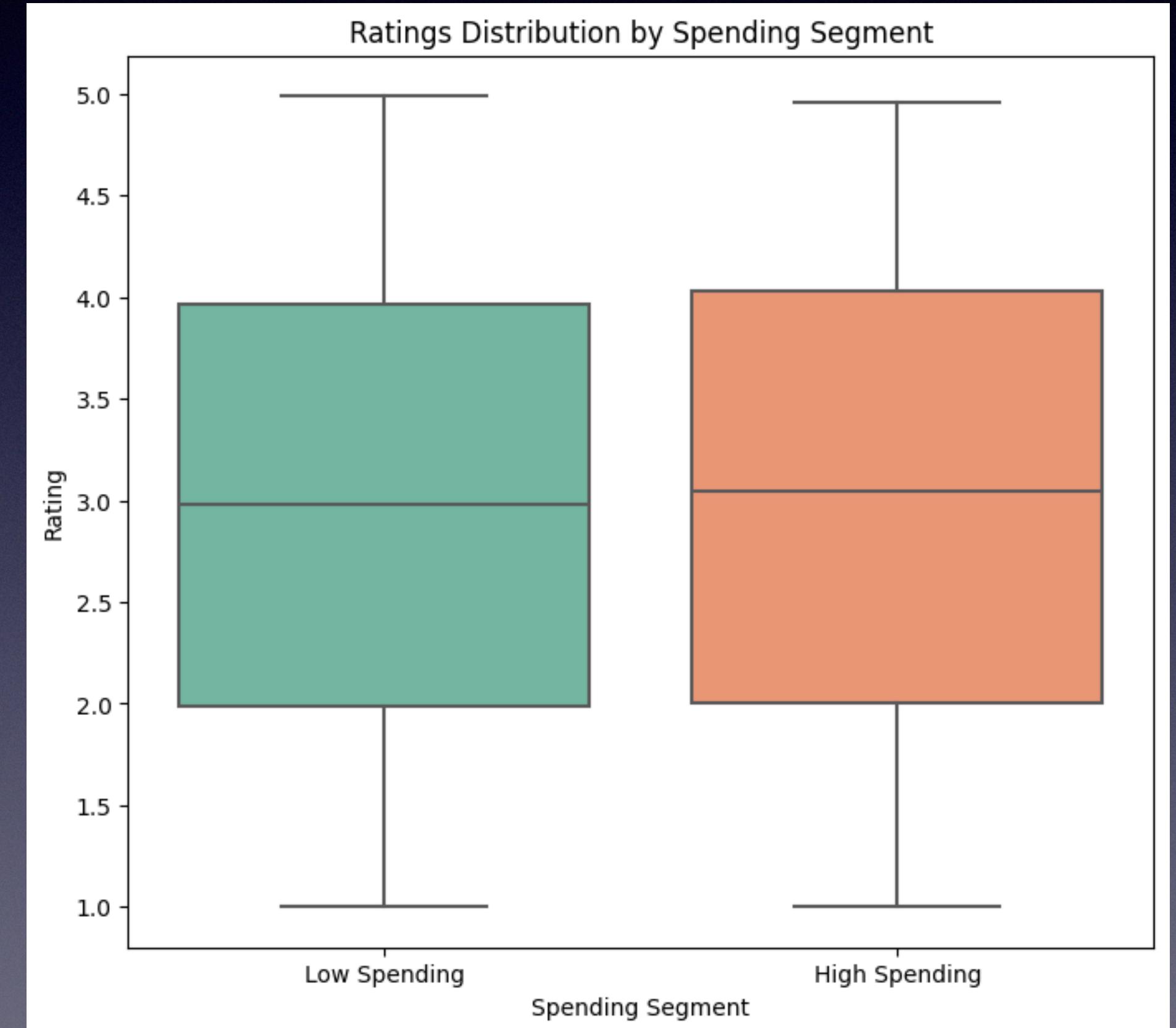


Distribution of Ratings per Segment

- **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	3.04
Low - Spending	2.97



Insight: High spenders tends to rate products slightly more positively than low spenders.

Recommendation and Key Findings

- 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.

