EDA REPORT

Prepared By:

Aditya Shinde

Table Of Contents

Generate product catalog insights on category/subcategory level.

As companies in the 21st century keep making their operations digital, the importance of data in decision-making has become more and more crucial. One area where data is really valuable but often ignored is product catalog data. It contains a lot of useful information that can help make smart decisions. Specifically, insights about different categories and subcategories can give us a detailed understanding of things like what products we have in stock, what customers like, and how well our sales are doing.

Understanding Trends and Preferences

A logical starting point is to understand customer preferences. By analyzing the gender, occasion, jewelry type, metal, metal color, gold karat, and diamond clarity, this can uncover popular trends across categories and subcategories. This can inform future designs, marketing strategies, and sales forecasts. For instance, a trend may show that women prefer white gold engagement rings with high diamond clarity. This will be explained in more detail in a future report.

Assessing Price and Discounting Strategies

Factors like price, discounted price, and discount percentage can be invaluable for gleaning insights about pricing strategies. Analyzing these aspects at the category and subcategory levels can reveal how pricing and discounts affect the popularity of items. It might be the case that certain types of jewelry or specific collections are more sensitive to price adjustments than others.

Stock Management and Availability

Availability and manufactured date information enable tracking the inventory changes over time. For instance, certain subcategories may have products that sell out quickly, requiring more frequent restocking. Trends in these areas can also give insights into seasonal demand fluctuations, informing better inventory management practices.

Leveraging Search and Trendrank

Google product category and trend rank can provide valuable information about the popularity of categories and subcategories online. They can reveal which products are more frequently searched or trending, leading to improved search engine optimization strategies and better alignment with consumer demand.

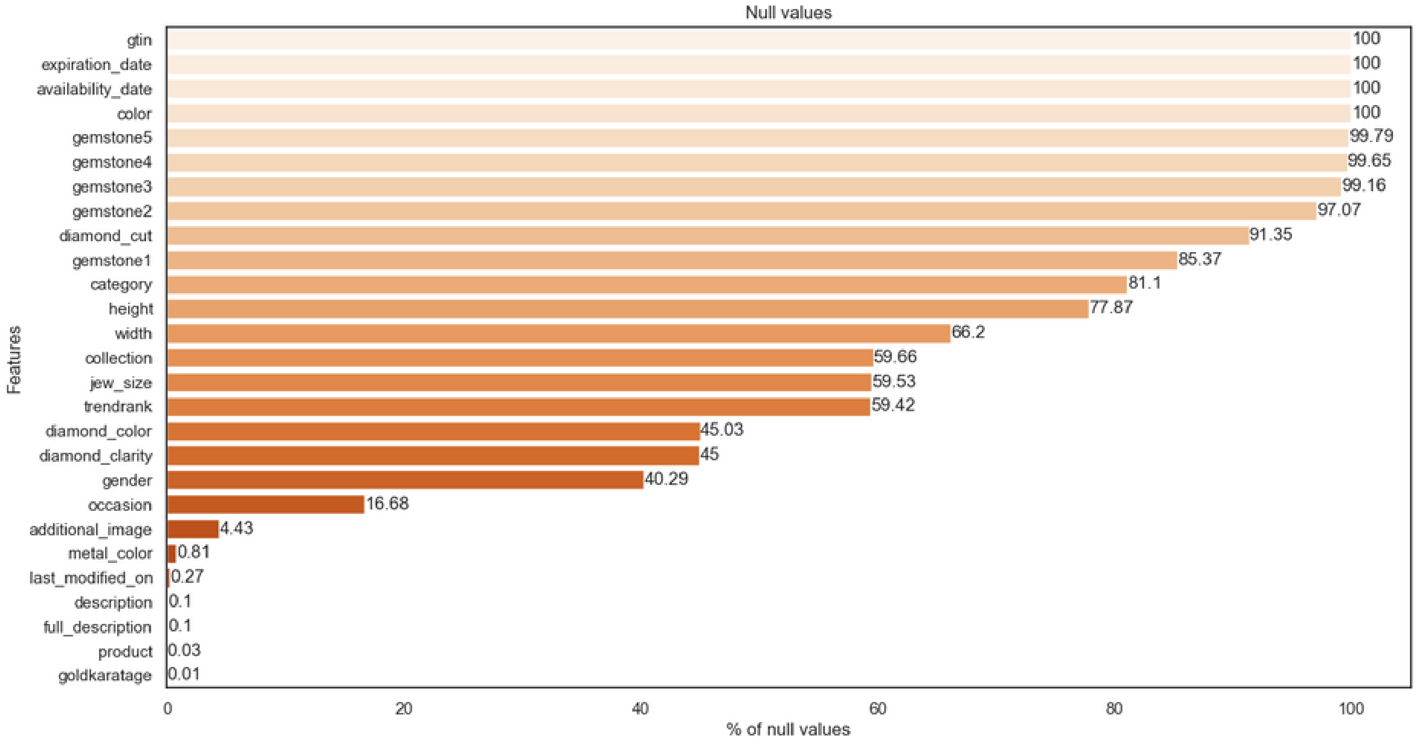
Data Preprocessing: The first phase of the analysis involved an in-depth data cleaning and preprocessing exercise, crucial for the quality and reliability of the subsequent insights.

Step 1: Initiated the process by importing the necessary libraries required for the analysis and loaded the provided dataset.

Step 2: An initial investigation of the data dimensions was conducted to understand the scope of the data. This helped us identify the volume of the data and the number of attributes under analysis.

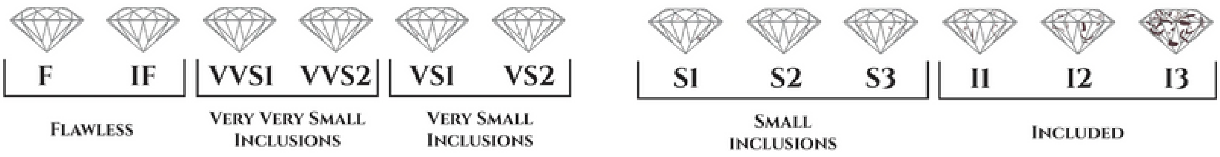
Step 3: To focus our analysis on features with significant variability, identified and removed all columns whose variance was 1 or maximized. This helps to limit the effect of redundant features and prevent overfitting in any subsequent model.

Step 4: Explored the data to identify any null values. visualized the distribution of these values using a graphical plot, which gave us a clear understanding of the data's completeness. Following this, made an informed decision on which columns to fill or remove based on the extent of the missing data and the column's significance.



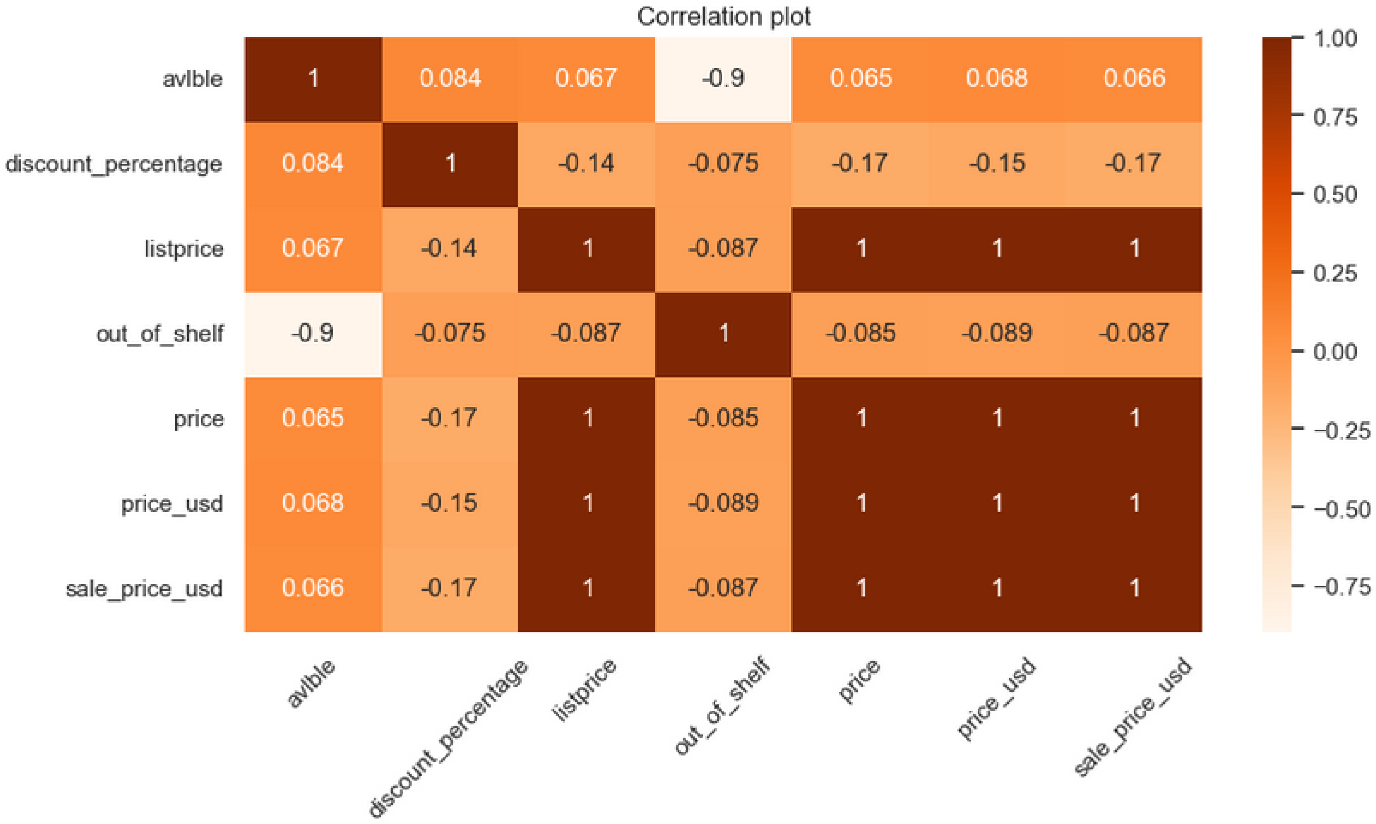
Based on the null value percentage, dropped columns that contained a significant number of missing values because filling in the missing data would have resulted in introducing synthetic or fabricated information. filled the missing values with an "unknown" category, acknowledging that the specific information is unavailable, while features with less than 1% missing values were filled using the mode method. This ensured proper handling of missing data without compromising the dataset's integrity.

Step 5: Next, performed feature engineering on several columns to make the data more meaningful and easier to analyze. Then converted the Unix Timestamp format column into a standard DateTime format, removed unnecessary types, and consolidated the diamond clarity attribute into a few meaningful categories.



The image above assisted in gathering and consolidating information about the clarity of diamonds.

Step 6: A correlation plot was generated to better understand the interdependencies between various attributes in the dataset.



	avlble	discount_percentage	listprice	out_of_shelf	price	price_usd	sale_price_usd
avlble	1	0.084	0.067	-0.9	0.065	0.068	0.066
discount_percentage	0.084	1	-0.14	-0.075	-0.17	-0.15	-0.17
listprice	0.067	-0.14	1	-0.087	1	1	1
out_of_shelf	-0.9	-0.075	-0.087	1	-0.085	-0.089	-0.087
price	0.065	-0.17	1	-0.085	1	1	1
price_usd	0.068	-0.15	1	-0.089	1	1	1
sale_price_usd	0.066	-0.17	1	-0.087	1	1	1

This correlation plot is to identify and eliminate columns that exhibit a high correlation with each other, regardless of whether the correlation is positive or negative. Since these columns provide redundant information, By doing so, can reduce the dimensionality of the dataset without losing significant information. This dimensionality reduction enhances computational efficiency and improves the interpretability of the model.

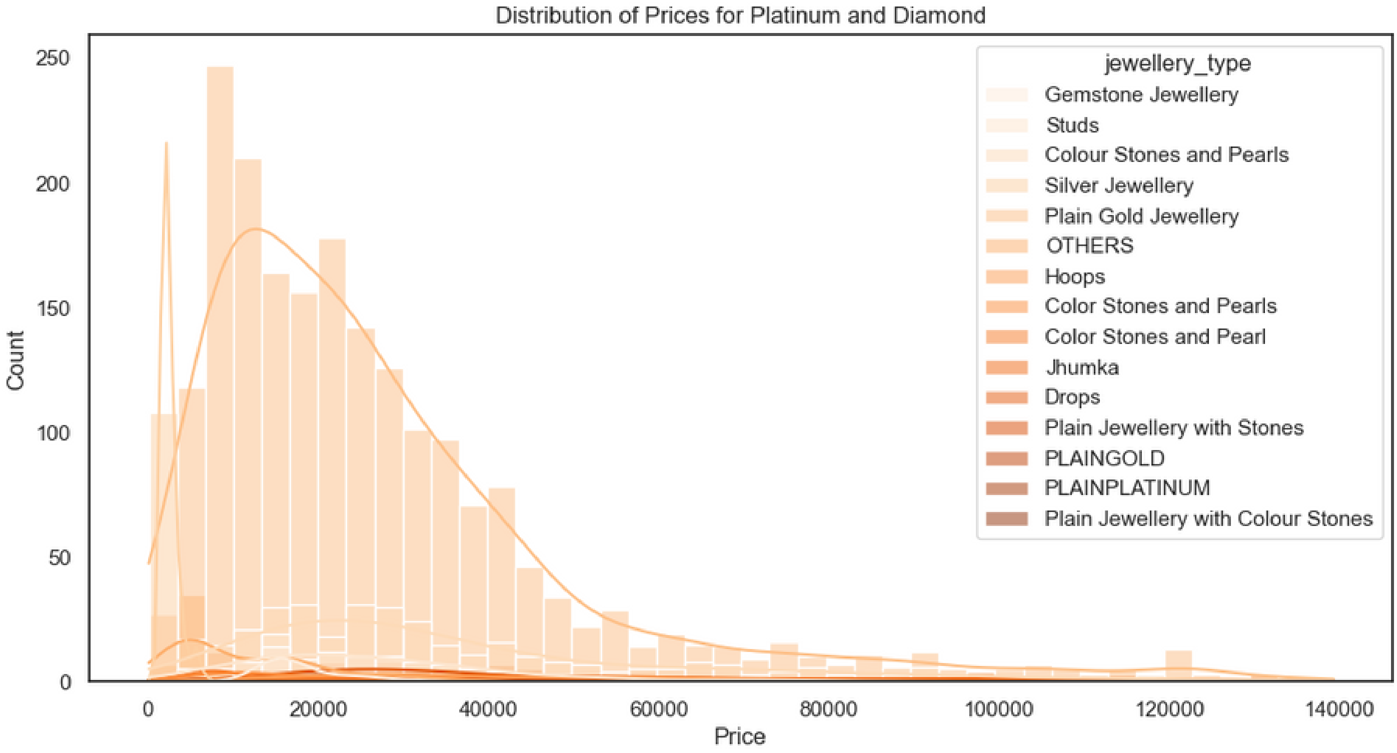
Here avlble & out_of_shelf are negatively correlated, and listprice, price, price_usd, and sale_price_usd are perfectly correlated. hence right action has been taken on one of these columns

Step 7: Finally, to ensure data consistency and improve the analysis's accuracy, then converted all columns to their appropriate data types.

Exploratory Data Analysis (EDA): After preprocessing the data, proceeded with the exploratory data analysis (EDA) to investigate the data's underlying patterns, spot anomalies, and check assumptions.

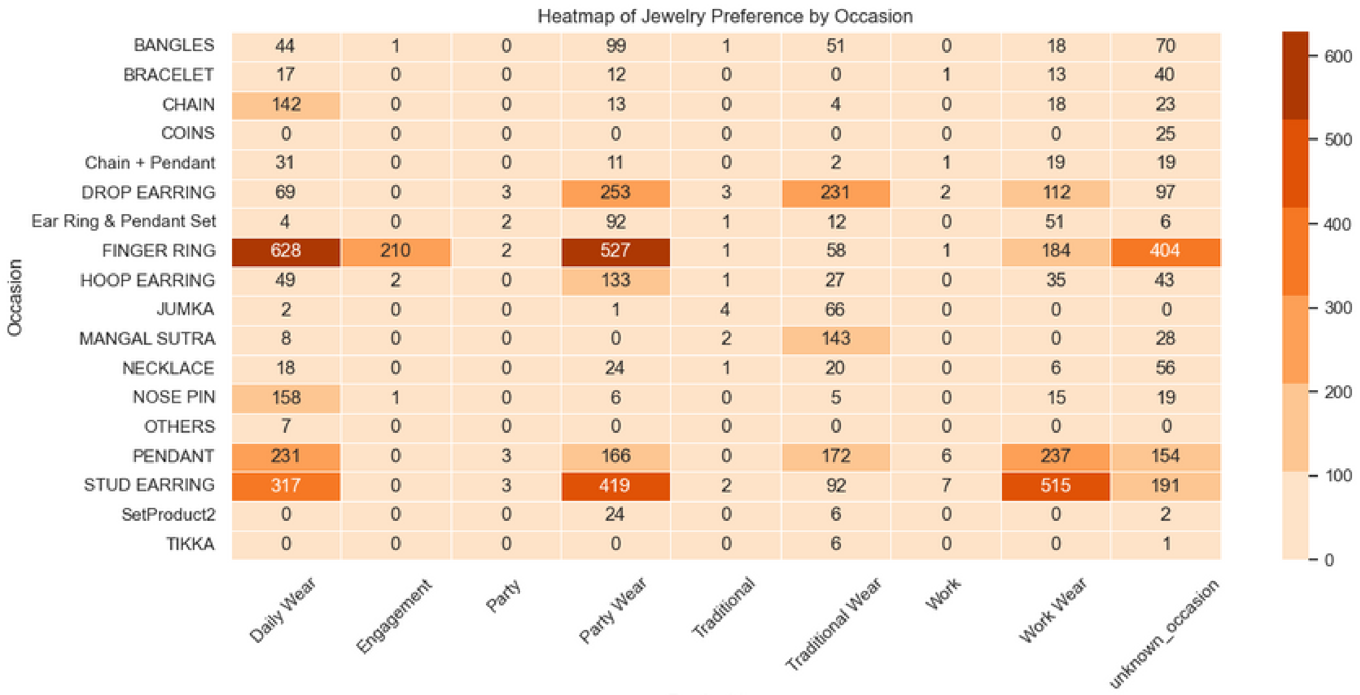
Describing the dataset in its entirety, providing a statistical summary of all the attributes.

Plot 1: A distribution plot of the product prices was created to visualize their range and dispersion.



Insight: This plot represents the price distribution, showing that the majority of transactions occur at a price of 20,000 rupees.

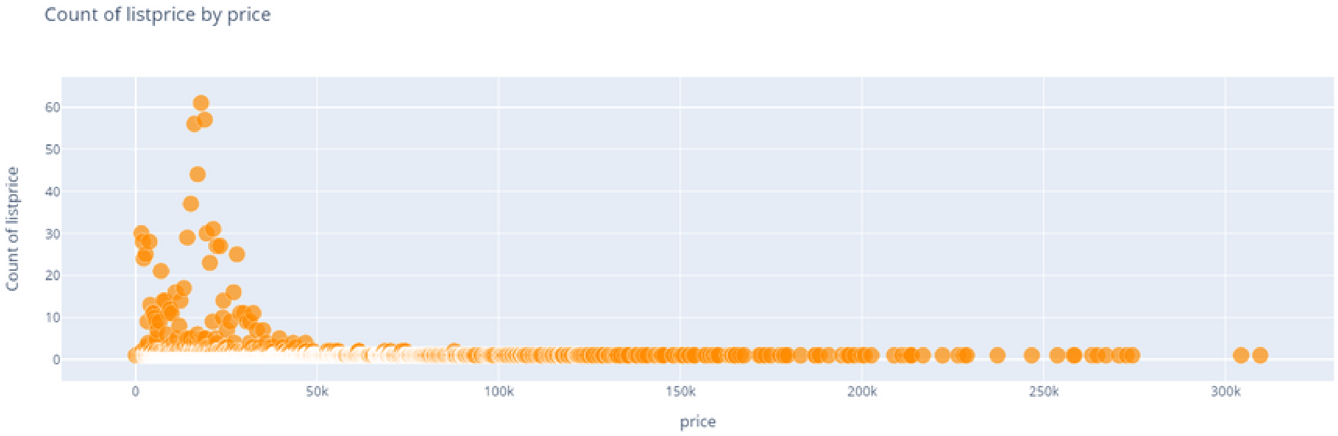
Plot 2: Generated a heatmap to identify trends and patterns in the relationship between different occasions and products.

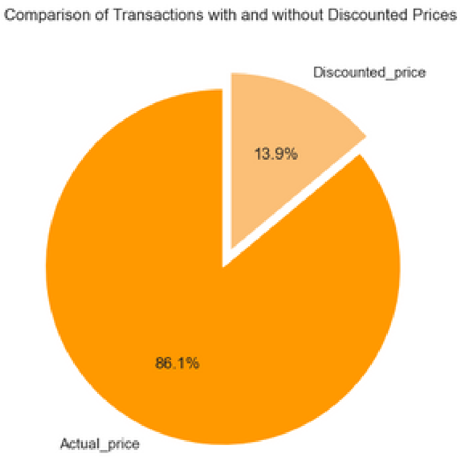


	Daily Wear	Engagement	Party	Party Wear	Traditional	Traditional Wear	Work	Work Wear	unknown_occasion
BANGLES	44	1	0	99	1	51	0	18	70
BRACELET	17	0	0	12	0	0	1	13	40
CHAIN	142	0	0	13	0	4	0	18	23
COINS	0	0	0	0	0	0	0	0	25
Chain + Pendant	31	0	0	11	0	2	1	19	19
DROP EARRING	69	0	3	253	3	231	2	112	97
Ear Ring & Pendant Set	4	0	2	92	1	12	0	51	6
FINGER RING	628	210	2	527	1	58	1	184	404
HOOP EARRING	49	2	0	133	1	27	0	35	43
JUMKA	2	0	0	1	4	66	0	0	0
MANGAL SUTRA	8	0	0	0	2	143	0	0	28
NECKLACE	18	0	0	24	1	20	0	6	56
NOSE PIN	158	1	0	6	0	5	0	15	19
OTHERS	7	0	0	0	0	0	0	0	0
PENDANT	231	0	3	166	0	172	6	237	154
STUD EARRING	317	0	3	419	2	92	7	515	191
SetProduct2	0	0	0	24	0	6	0	0	2
TIKKA	0	0	0	0	0	6	0	0	1

Insight: This heatmap indicates that the majority of sales are occurring for finger rings, particularly for daily wear and party wear, with earrings and pendants following as the next best-selling items.

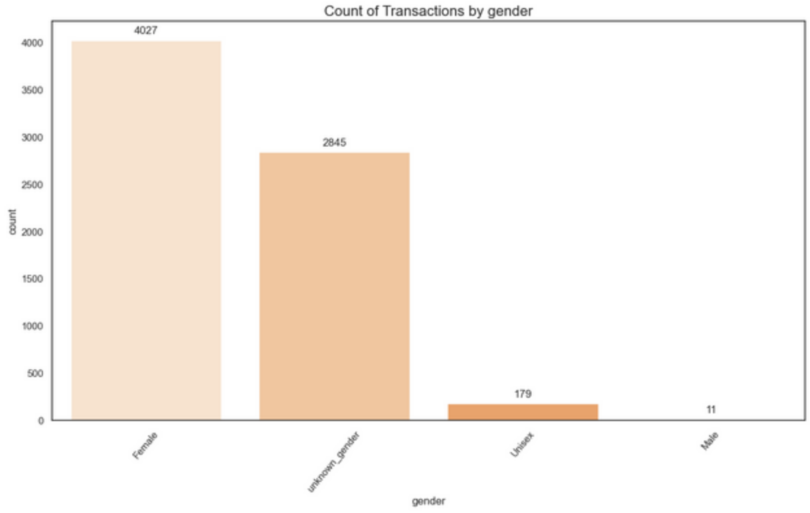
Plot 3: A scatter plot was created to compare the relationship between the price and discounted price, providing insights into the discount strategies across the catalog. followed by an exploding pie chart that was used to visually represent the transactions made with a discounted price. This graphical representation helps to quickly identify the proportion of discounted sales.

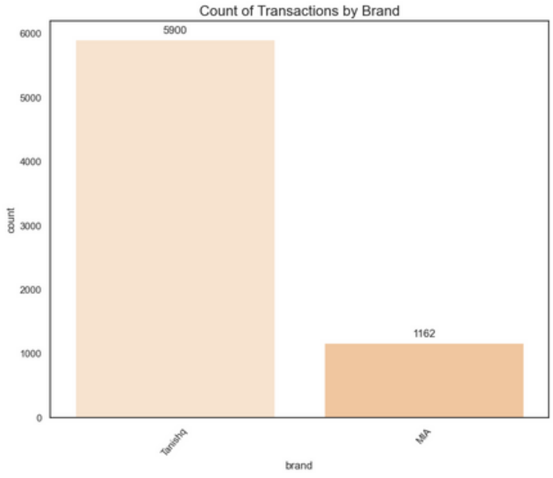




Insight: The scatter plot indicates that the majority of discounts were applied to products priced below 50000rs, while the pie chart shows that only 13.9% of transactions received a discounted price.

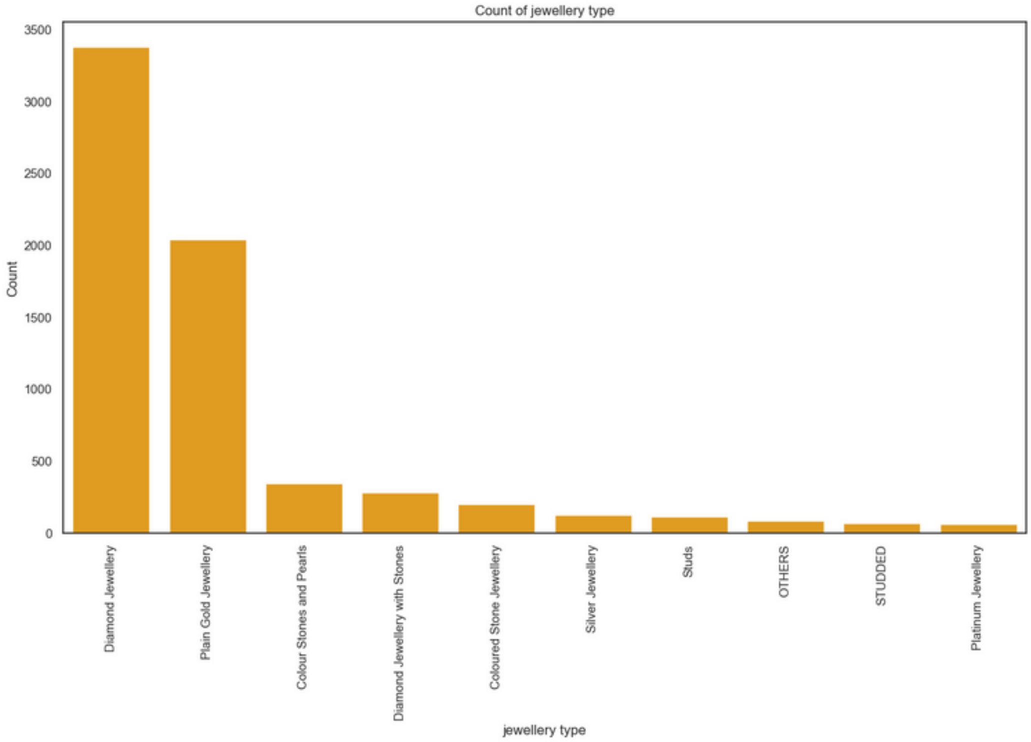
Plot 4: To understand the distribution by gender and brand, plotted a count plot.





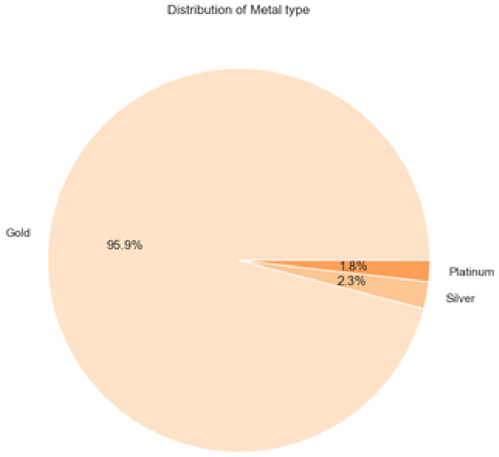
Insight: Although there have been only 11 transactions made by customers in the male category, it is not reliable to draw conclusions based on this information alone because there is also an unknown type of customer. Additionally, the second plot indicates that Mia, a new brand from Titan, has generated relatively lower sales.

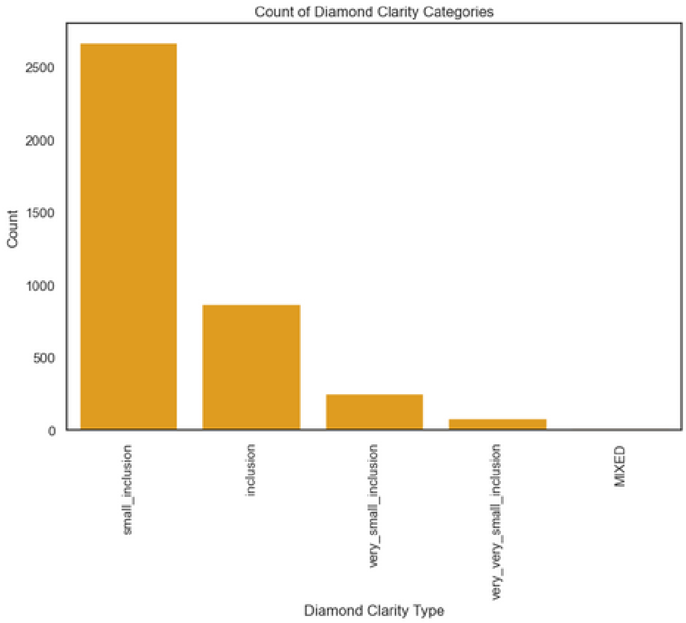
Plot 4: Plotted a bar graph showcasing jewelry types to understand product preference and trends in market.



Insight: This plot indicates that there is a higher demand for Diamond jewellery and gold jewellery, suggesting that greater attention should be given to manufacturing and designing of these kinds of jewellery.

Plot 5: Lastly, a pie chart was created to visualize the distribution of various metal types in our catalog, providing insights into the most popular types and a bar graph showcasing diamond clarity types.





Insight: This data reveals that over 95% of individuals have a preference for purchasing gold, according to the first plot. Additionally, the second plot demonstrates that sales of diamonds with minor impurities, specifically those of the small inclusion type, surpass those of other diamond types.

Conclusion: This comprehensive analysis of our product catalog has led to a deeper understanding of our offerings, consumer preferences, and price dynamics. The insights gleaned from this exercise will prove invaluable for making data-driven decisions and further tailoring our catalog to better meet our customer's needs. Ultimately, businesses can use these insights to make informed decisions that boost sales, enhance customer satisfaction, and drive growth.

Segmentation of Users Based on Browsing Activity

Introduction

Understanding user behavior is important in today's digital landscape. By segmenting users based on their browsing activities, organizations can gain actionable insights to optimize their platforms, provide personalized experiences, and thus foster customer loyalty. In this analysis, the journey to uncover the latent groups within our user base using the following features: action_type, object_type, user_type, epoch (timestamp), referrer, city/state, country, latitude, longitude, platform, browser, day_of_year, datetime and hour.

This segmentation enhances user engagement and satisfaction by presenting relevant information, thereby driving conversions and boosting sales. Furthermore, it aids in predicting future behaviors and identifying potential growth opportunities. This strategy, therefore, not only optimizes user experience but also enhances business performance, leading to a competitive advantage.

This analysis of this section outlines the methodology and findings of our user segmentation analysis. The data utilized was the event data set with more than 10 million records, capturing extensive browsing information, with the purpose of understanding user behavior and tailoring our approach to meet their needs more effectively.

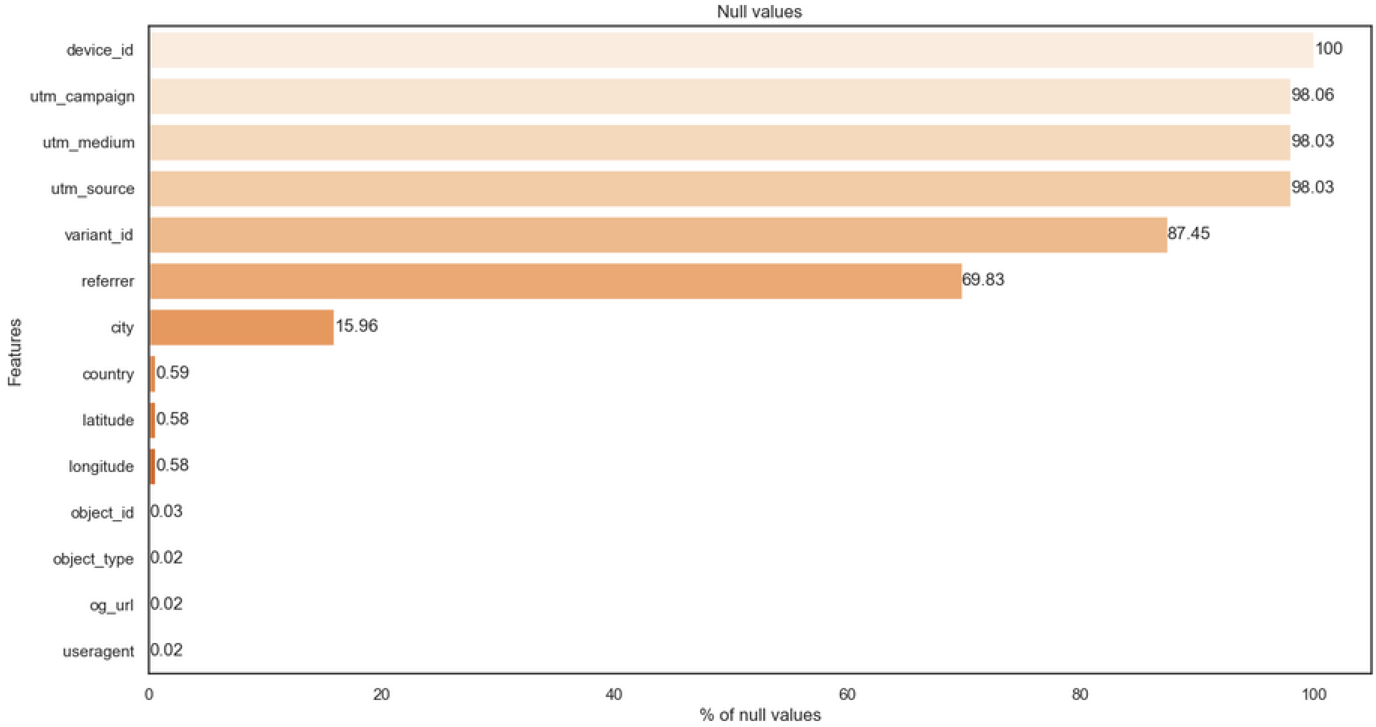
Data Preprocessing: The first phase of the analysis involved an in-depth data cleaning and preprocessing exercise, crucial for the quality and reliability of the subsequent insights.

Step 1: first step involved loading the browsing data using the necessary libraries. It's vital to ensure that the data is properly ingested and ready for subsequent analysis. Given the massive size of our dataset, appropriate resource allocation was critical.

Step 2: verified the dimensions of our dataset to confirm the breadth and depth of our data. This allows us to scale our analysis and the complexity of the potential relationships within the data.

Step 3: To avoid dimensional complexity, removed all columns with a variance of 1 or maximum. This step helped us focus on the more meaningful data while reducing redundancy and potential overfitting during the model-building phase.

Step 4: Checked for null values and removed features with significant null percentages.



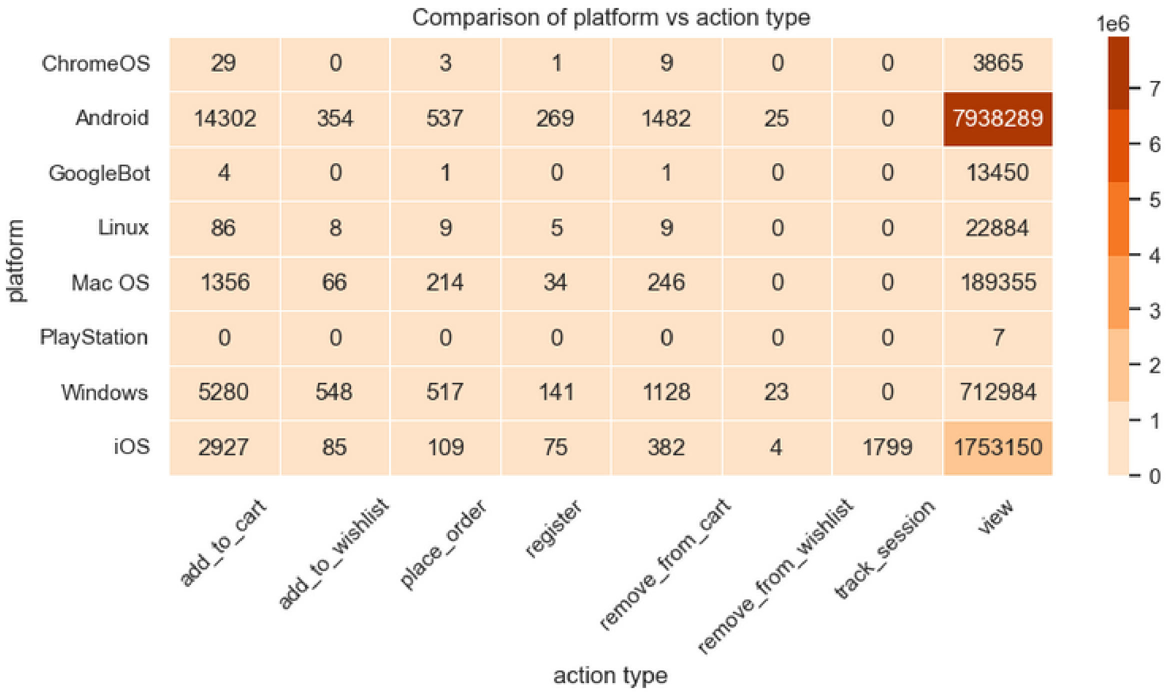
A notable exception was the 'city' column, which held around 16% null values; imputed these missing values using the 'geopy' library which extracts information from the latitude and longitude columns. The remaining null values were addressed by applying the mode method or substituting with an alternative category, ensuring a more reliable and robust dataset.

Step 5: Extracted valuable data from the 'user agents' column through feature engineering. By using the 'httpagentparser' library, determined the browser and platform information of each user, a crucial step in understanding user behaviour patterns and preferences.

Step 6: Another vital feature engineering step involved the datetime column, extracting relevant data such as dates and days. This provided deeper insights into temporal browsing patterns, contributing to our understanding of user activity timeframes.

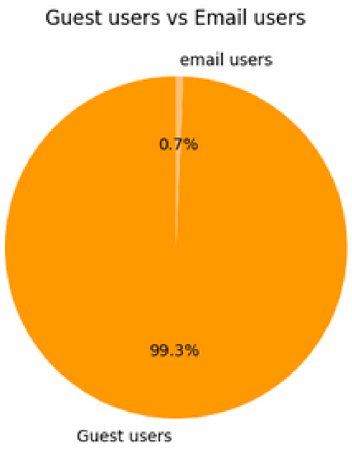
Exploratory Data Analysis (EDA): Following the preprocessing of the data, the next step involved conducting exploratory data analysis (EDA) to examine the inherent patterns in the data, identify any unusual observations, and validate assumptions.

Plot 1: The heat map plot provided insights into platform vs action type, depicting user behavior across different platforms. This visual aid helped us comprehend the correlation and possible interactions between different user actions and their chosen platforms.



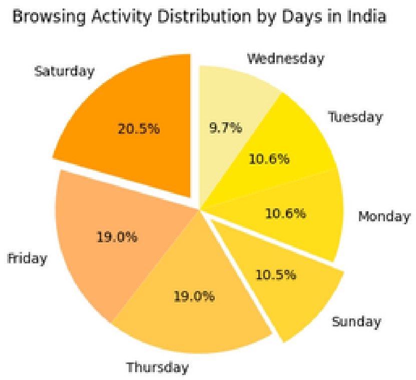
Insight: According to the data, the majority of users are utilizing the Android platform, indicating a preference for mobile phones. In comparison, Windows users are more inclined to add items to their cart and proceed with placing orders.

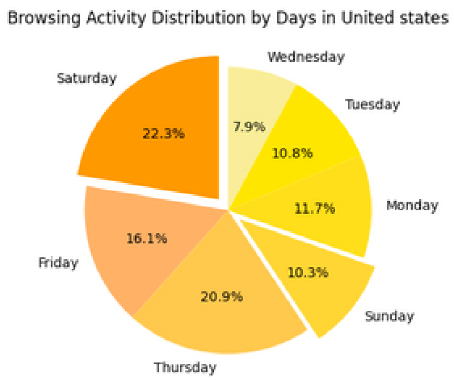
Plot 2: The pie chart offered a holistic view of the types of users viewing the website, providing us with a demographic breakdown and helping us understand the composition of our user base.



Insight: this plot lacks significant insights, but it does indicate the percentage of browsing activity among guest users in comparison to registered email users.

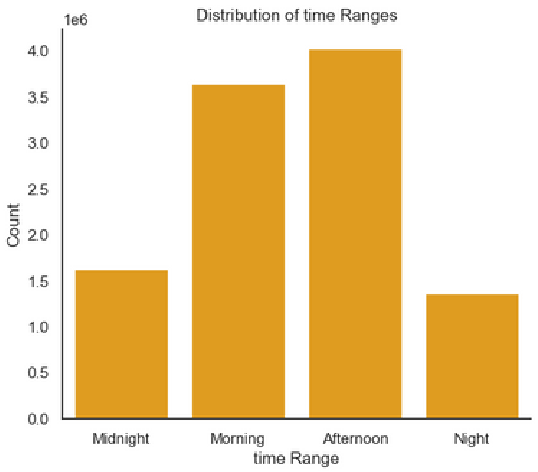
Plot 3: With an exploding pie chart, displayed the browsing activity distribution in India and the United States. This visualization highlighted the comparative engagement levels and browsing habits in these two countries, critical in tailoring regional strategies.

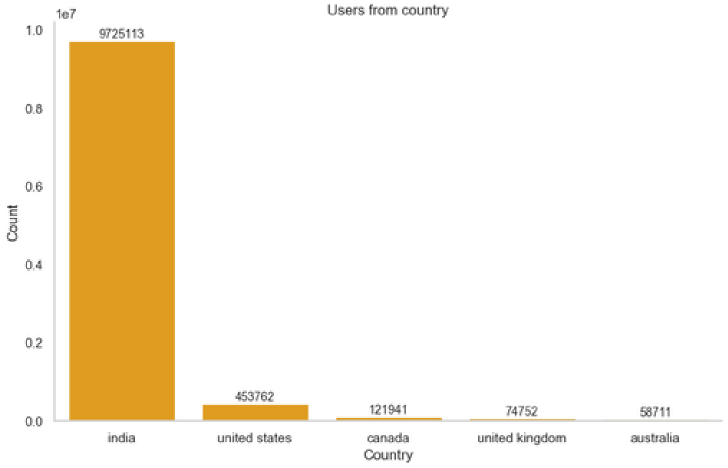




Insight: This plot demonstrates the comparison between Indian and United States users, revealing that both groups exhibit similar patterns of activity. It also reveals that people tend to utilize the website more on Saturdays and Fridays, while their usage is lowest on Wednesdays. This finding provides valuable insight for optimizing the distribution of website traffic and load balancing , thereby reducing server and computing costs.

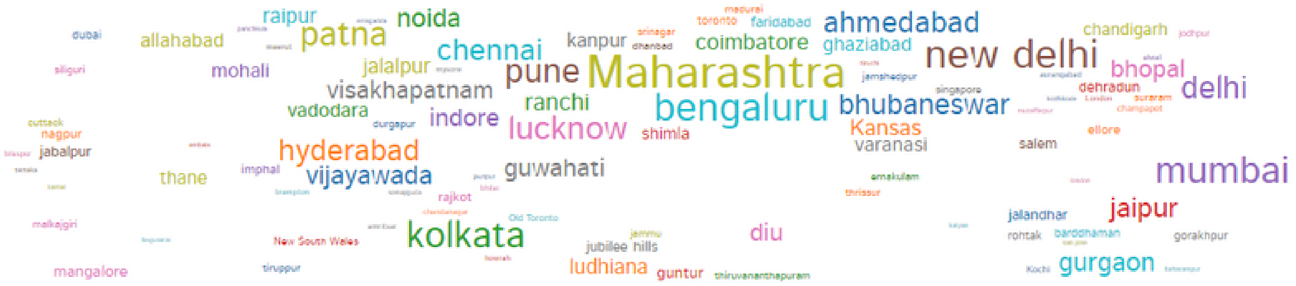
Plot 4: By using a bar chart, discerned the time of the day users are most active, equipping us to optimize website operations and communication strategies in alignment with peak user activity times, a second chart demonstrated which countries boasted higher user engagement with our website.





Insight: The first plot indicates that people tend to engage in browsing activities more frequently during the afternoon, which includes data from various time zones, resulting in slightly higher usage even during midnight hours. The second plot highlights that a significant portion of the website's traffic originates from users in the Indian region. This information is invaluable in assessing the global reach of the platform

Plot 5: The word cloud reveals the most frequent city or state users were browsing from, offering insights into geographic user distribution, which can further guide location-specific marketing and optimization strategies.



Insight: According to the word cloud plot, it is evident that the state of Maharashtra has the highest number of users, followed by New Delhi and Bengaluru. This data encompasses cities and states from around the world.

In conclusion, data preprocessing and analysis have yielded profound insights into our users' browsing behaviors and preferences. By segmenting users based on their browsing activity, we're able to develop tailored strategies to enhance user experience and engagement. The ongoing application of such data-driven methodologies will continue to be pivotal in our platform's optimization and growth.

Identifying the major factors affecting user's buying decisions

Introduction

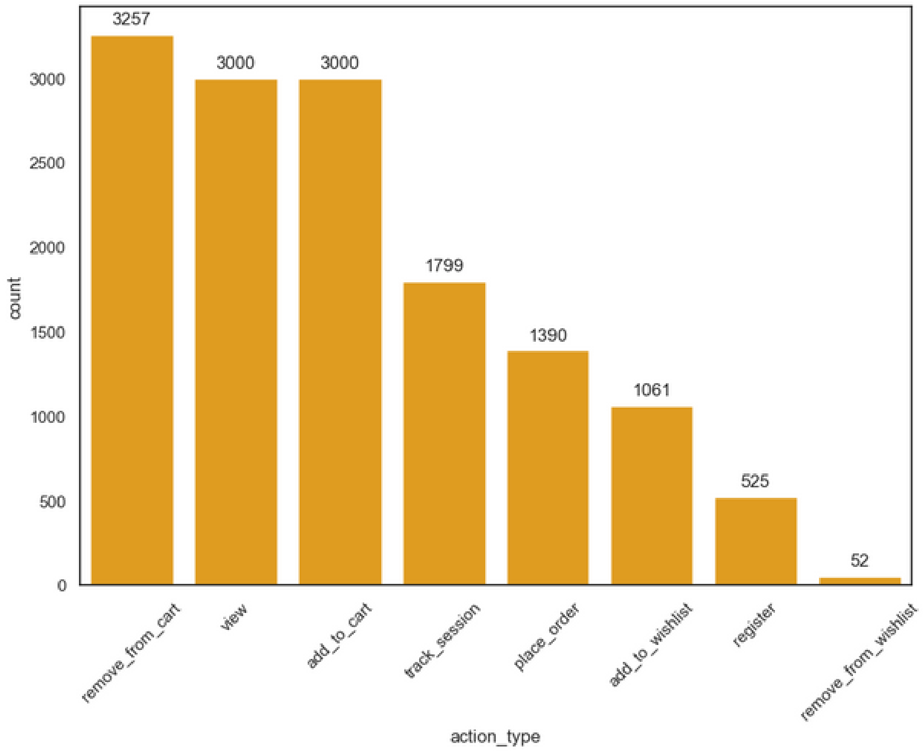
The objective was to identify the major factors affecting user's buying decisions. To achieve this, focused on two main features: 'Action Type' and 'Time Duration'. Each attribute was used as a target column to construct two different machine learning models using XGBoost, one as a multiclassification problem and the other as a regression problem.

Model Building with 'Action Type' as Target

Step 1: Started by importing the necessary libraries and loading the preprocessed dataset. To mitigate the issue of imbalanced data, executed careful feature engineering, ensuring all categories in 'Action Type' were proportionately represented. date information was extracted using datetime functions and data was properly formatted for further analysis.

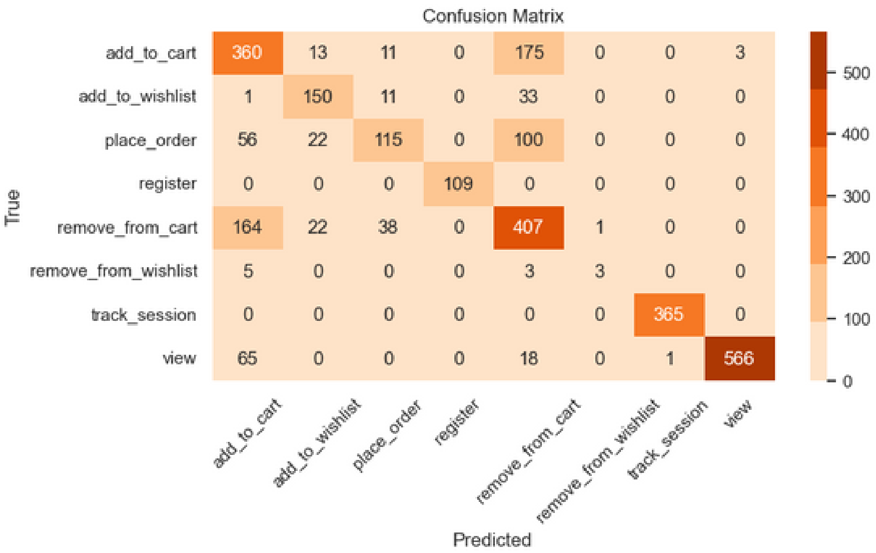
Step 2: For categorical features, weopted for label encoding instead of one-hot encoding to avoid dimensional complexity. Additionally, wenormalized all integer and float columns, reducing bias and helping the model perform more efficiently.

Step 3: To counter the imbalanced nature of our data, a random selection of records was used to balance the category counts.

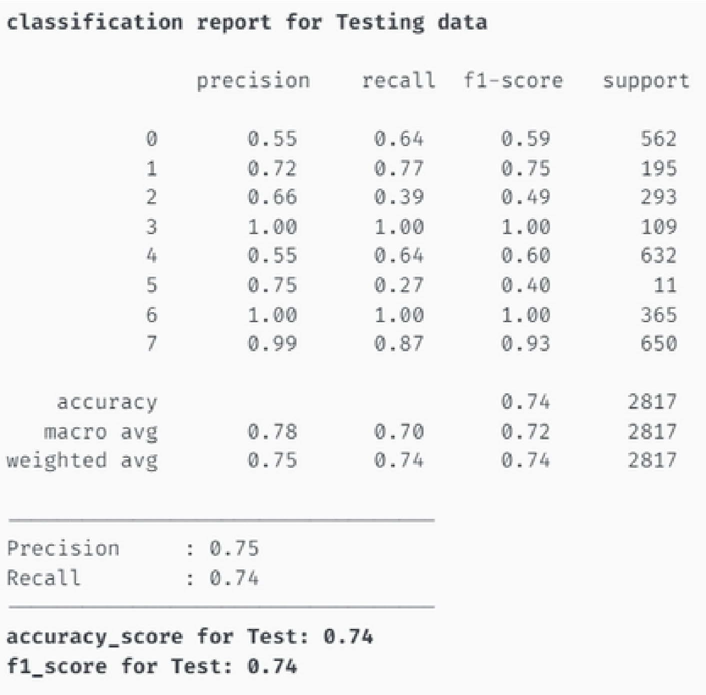


Step 4: The data was then split in an 80-20 proportion for training and testing purposes. The XGBoost algorithm, widely recognized for its performance in machine learning, was utilized for model training.

Step 5: Our model's performance was gauged using a confusion matrix and classification report. This allowed us to see a detailed breakdown of the model's ability to classify different actions users might take.



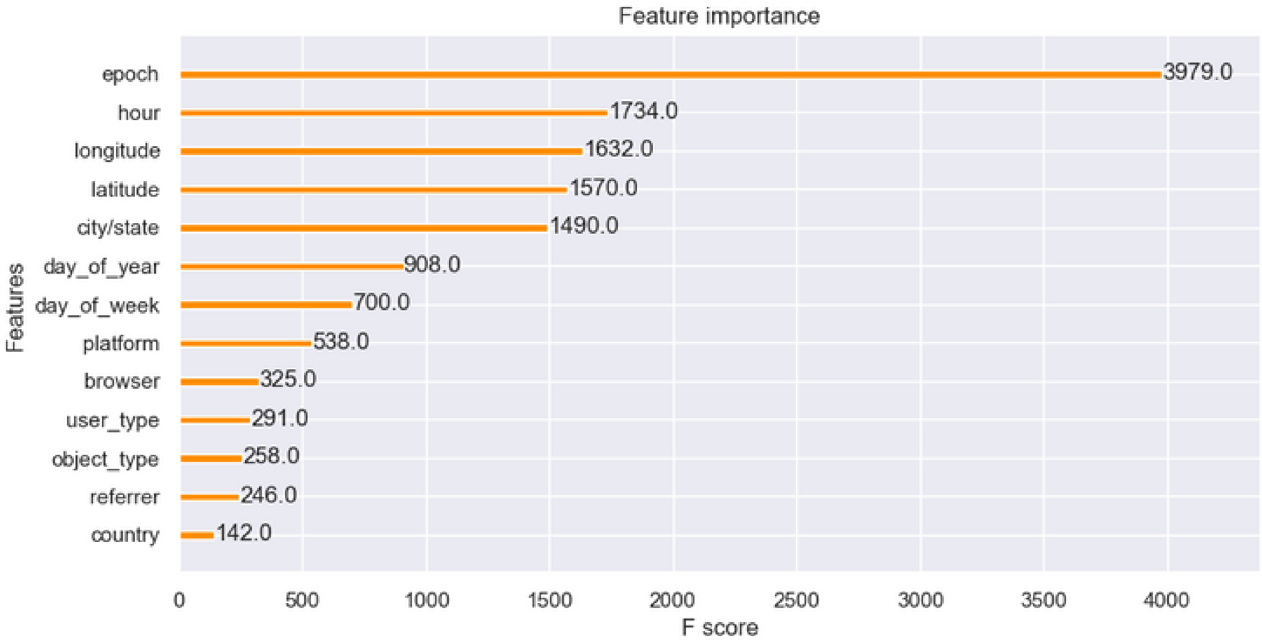
	add_to_cart	add_to_wishlist	place_order	register	remove_from_cart	remove_from_wishlist	track_session	view
add_to_cart	360	13	11	0	175	0	0	3
add_to_wishlist	1	150	11	0	33	0	0	0
place_order	56	22	115	0	100	0	0	0
register	0	0	0	109	0	0	0	0
remove_from_cart	164	22	38	0	407	1	0	0
remove_from_wishlist	5	0	0	0	3	3	0	0
track_session	0	0	0	0	0	0	365	0
view	65	0	0	0	18	0	1	566



```
classification report for Testing data

              precision    recall  f1-score   support

           0       0.55      0.64      0.59       562
           1       0.72      0.77      0.75       195
           2       0.66      0.39      0.49       293
           3       1.00      1.00      1.00       109
           4       0.55      0.64      0.60       632
           5       0.75      0.27      0.40        11
           6       1.00      1.00      1.00       365
           7       0.99      0.87      0.93       650

    accuracy                           0.74      2817
   macro avg       0.78      0.70      0.72      2817
weighted avg       0.75      0.74      0.74      2817

__________________________________
Precision      : 0.75
Recall         : 0.74
__________________________________
accuracy_score for Test: 0.74
f1_score for Test: 0.74
```

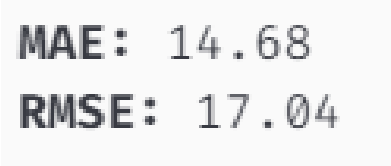
Insights of Model: Obtained a satisfactory score by utilizing a dataset of just 14,000 rows. It is likely that training the model with a larger dataset will yield even better performance. We can observe an accuracy and F1 score of 0.74, indicating a fairly accurate prediction. However, there are a few misclassifications specifically in the "place order" and "remove from cart" columns.

Step 5: Concluded the modeling process by plotting a feature importance graph. This graph provides crucial insights into the significant features that influence users' engagement and purchasing decisions.



Model Building with 'Time Duration' as Target

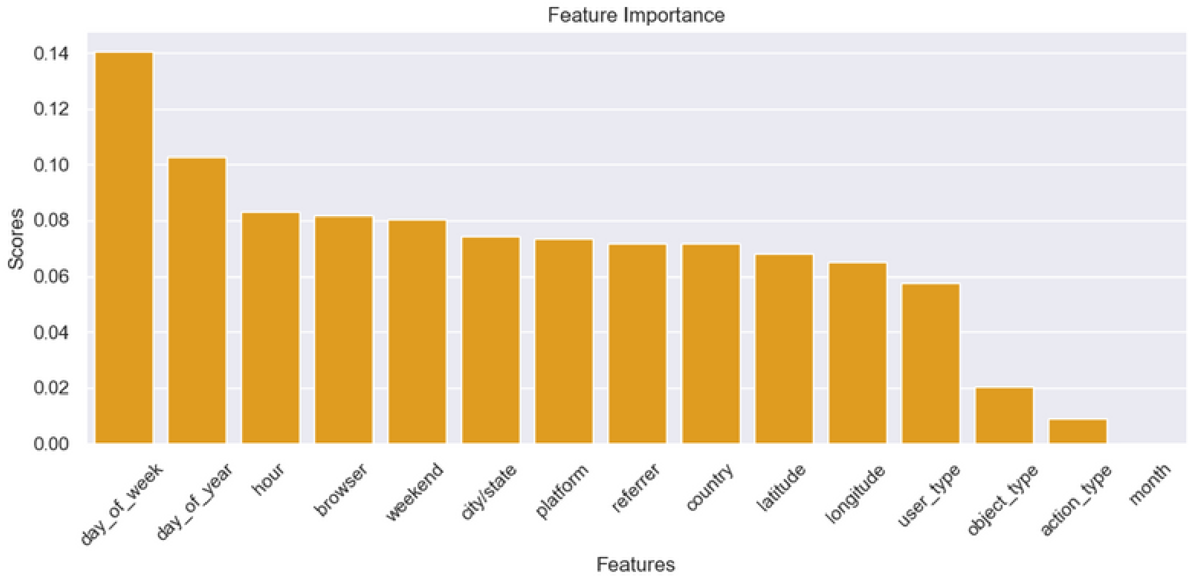
Step 1: the initial steps remain the same. First, preprocess the data by handling missing values, encoding categorical variables, and performing feature scaling if necessary. Next, split the dataset into training and testing sets. Then, define the XGB regressor model, Finally, fit the model to the training data, and evaluating on test data



```
MAE: 14.68
RMSE: 17.04
```

Insights of Model: The model exhibited a strong performance, indicating an accurate prediction of user's duration on the site with less error.

Step 2: Lastly, a feature importance graph was plotted. This graph underlined the significant features that can help us encourage users to stay longer on the site. The likelihood of someone staying longer and engaging in activities is influenced by the day of the week and the specific day of the year.



The comprehensive analysis and modeling conducted have provided us with valuable insights into the factors influencing users' buying decisions. these models can predict a user's actions and time spent on the website, allowing us to implement strategic changes and improvements to optimize user engagement and purchasing activities.