E-commerce Consumer Behavior Data Insights

Liu, Hai-Sung (Eason)

Dataset Background

The Online Shoppers Intention Dataset provides session-level data from an e-commerce website, including page views, durations, traffic sources, bounce/exit rates, and whether the user generated revenue. This dataset offers a unique opportunity to study online shopping behavior and purchasing decisions based on users' navigation patterns and session attributes.

Project Objective

The project aims to uncover behavioral trends and predictive features that influence consumer decision-making. It leverages data analysis tools such as Pandas for preprocessing, Matplotlib and Seaborn for multi-dimensional visualization, and pivot tables to summarize complex behavior hierarchies.

Data preprocessing

- Data Type and Structure Standardization

To maintain consistency, we ensured that each column had the appropriate data type. For example, boolean columns such as Weekend and Revenue were converted to categorical or numeric types where needed, facilitating aggregation and visualization later in the analysis pipeline

```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  category
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  category
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), category(2), float64(7), int64(7)
memory usage: 1.4 MB
```

- Handling missing and duplicated data

To ensure the integrity of the dataset, we first checked for missing values and duplicate records. Fortunately, the dataset was clean in this aspect, containing no null entries or duplicated rows. This eliminated the need for imputation or deduplication procedures, allowing us to proceed confidently with analysis

	0
Administrative	0
Administrative_Duration	0
Informational	0
Informational_Duration	0
ProductRelated	0
ProductRelated_Duration	0
BounceRates	0
ExitRates	0
PageValues	0
SpecialDay	0
Month	0
OperatingSystems	0
Browser	0
Region	0
TrafficType	0
VisitorType	0
Weekend	0
Revenue	0

dtype: int64

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
158	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0	0.0	Feb	1	1	1	3	Returning_Visitor	False	False
159	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0	0.0	Feb	3	2	3	3	Returning_Visitor	False	False
178	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0	0.0	Feb	3	2	3	3	Returning_Visitor	False	False
418	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0	0.0	Mar	1	1	1	1	Returning_Visitor	True	False
456	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0	0.0	Mar	2	2	4	1	Returning_Visitor	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11934	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0	0.0	Dec	1	1	1	2	New_Visitor	False	False
11938	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0	0.0	Dec	1	1	4	1	Returning_Visitor	True	False
12159	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0	0.0	Dec	1	1	1	3	Returning_Visitor	False	False
12180	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0	0.0	Dec	1	13	9	20	Returning_Visitor	False	False
12185	0	0.0	0	0.0	1	0.0	0.2	0.2	0.0	0.0	Dec	8	13	9	20	Other	False	False

125 rows × 18 columns

- Categorical variable cleaning

Several categorical variables required transformation for consistency and readability. We standardized the values of the VisitorType column by mapping inconsistent labels (e.g., Returning_Visitor, New_Visitor) to more concise categories. For the Month column, we reordered the months based on their natural calendar sequence to prevent disorder in time series visualization.

Month	count
Feb	181
Mar	1860
May	3329
June	285
Jul	432
Aug	433
Sep	448
Oct	549
Nov	2982
Dec	1706

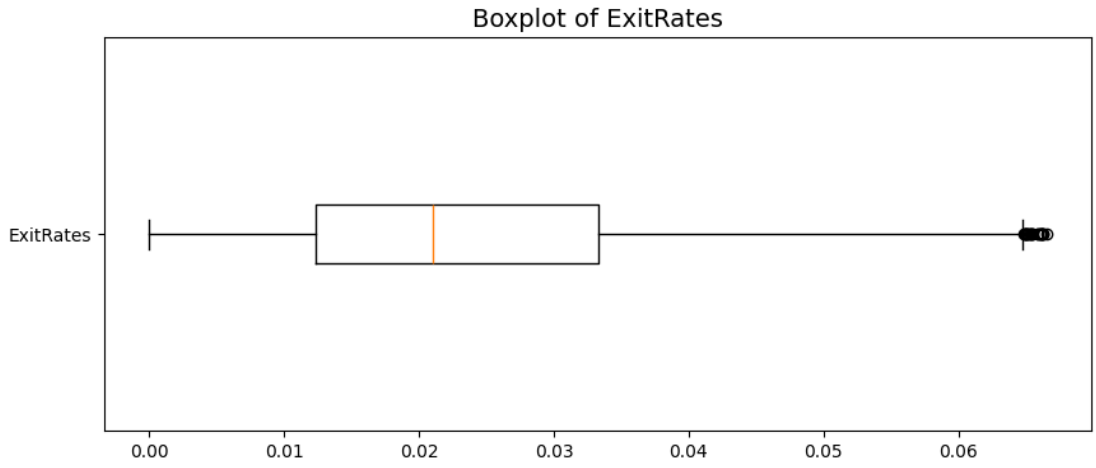
dtype: int64

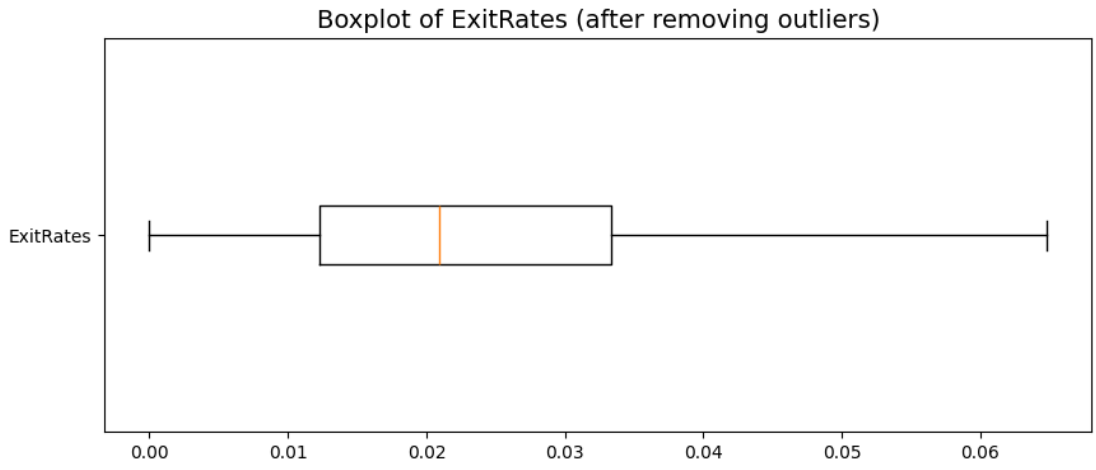
VisitorType	count
New_Visitor	1693
Other	81
Returning_Visitor	10431

dtype: int64

- Outlier Detection and Removal

Outlier detection was conducted using the Interquartile Range (IQR) method for continuous variables such as ExitRates. Data points beyond 1.5 times the IQR were removed to reduce skewness and enhance the robustness of downstream analysis. BounceRates, although showing long tails, was retained for its analytical significance.

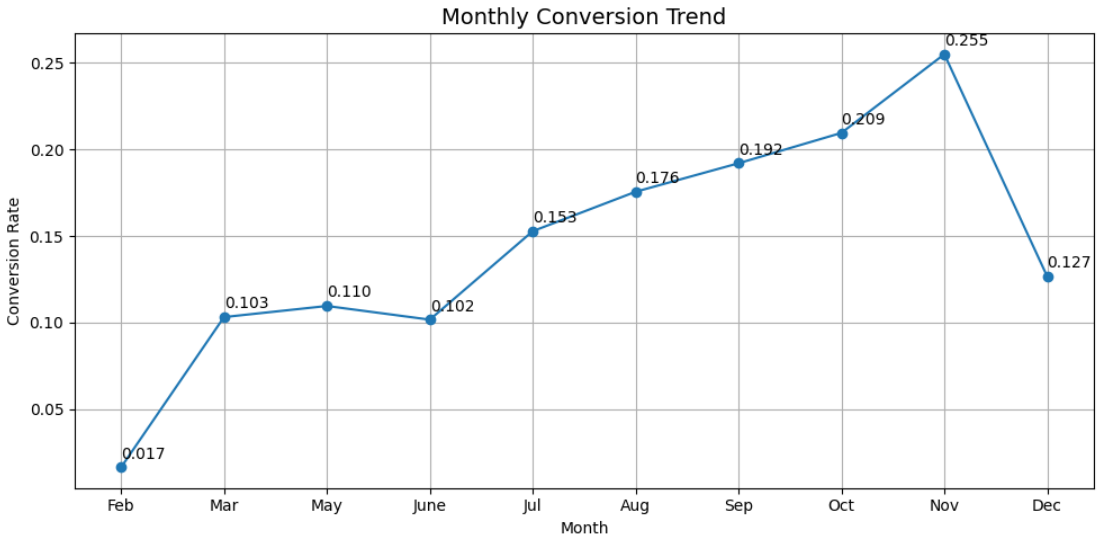




Exploratory Data Analysis

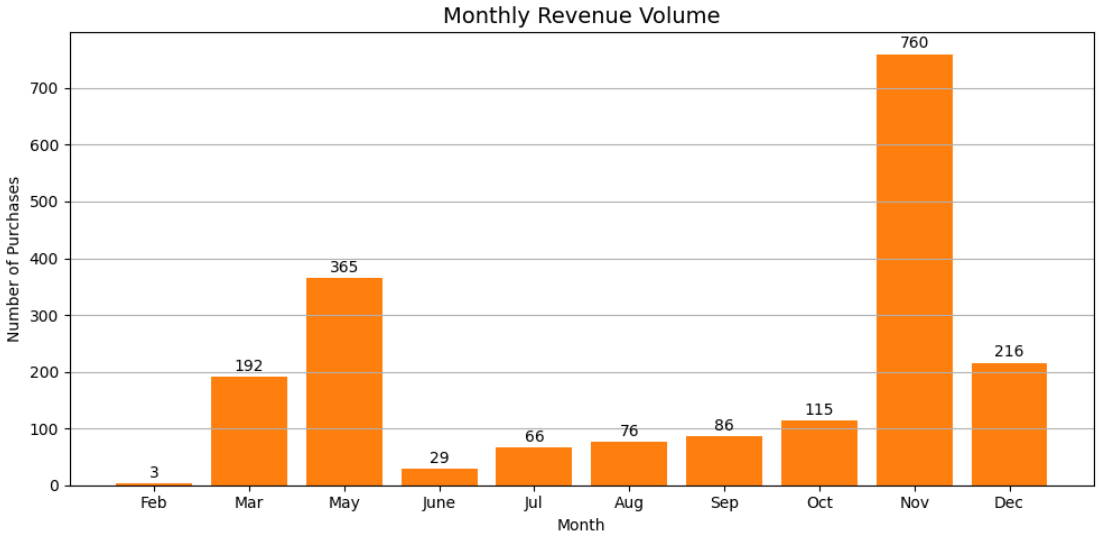
1. Monthly Trend Analysis
    - Average conversion rate by month

We first examined how the average conversion rate varies across different months. The line chart shows that conversion rates tend to peak around May and November, possibly corresponding to seasonal promotions. This insight reveals user purchasing behavior trends over time.



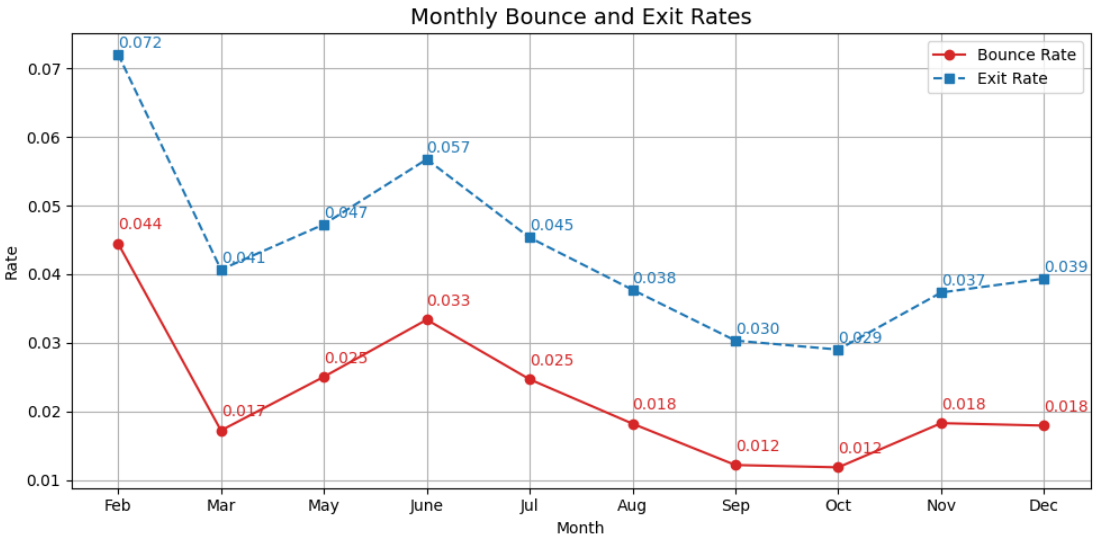
- Purchase volume by month

We then investigated the number of purchases (Revenue=True) across months. November shows the highest transaction count, highlighting peak shopping behavior that may coincide with global shopping festivals.



- Exit/Bounce rate change by month

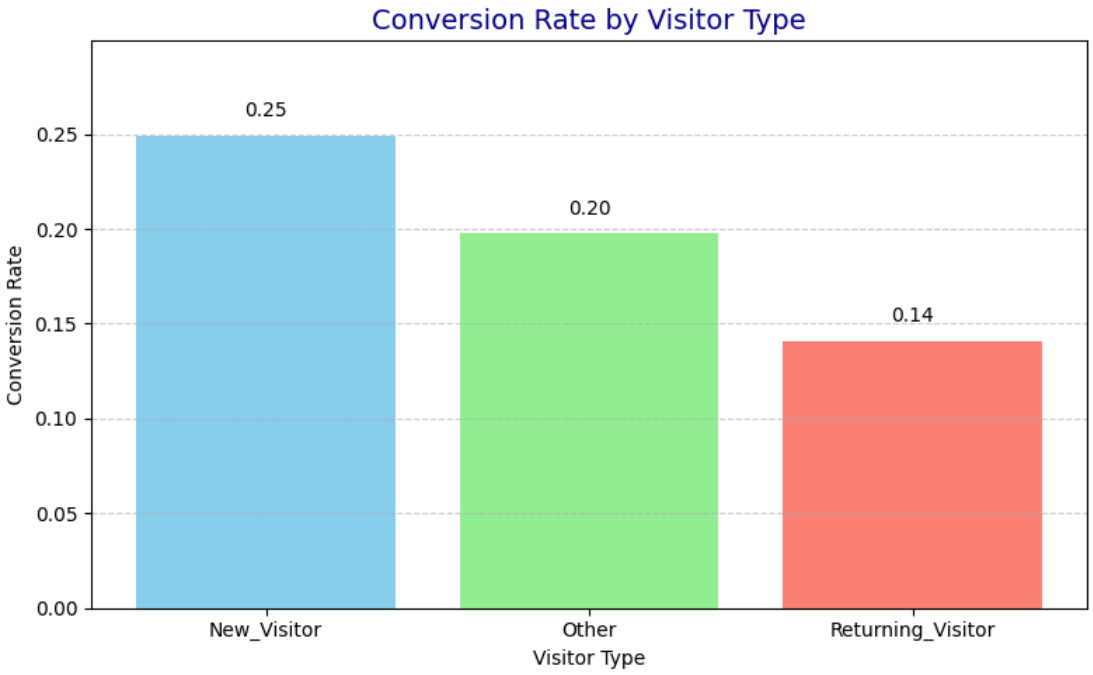
We further explored user behavior by tracking average Exit and Bounce rates by month. The relatively higher values in early months suggest more frequent user abandonment, which may help platforms optimize early-stage engagement strategies.



2. Visitor Conversion Analysis

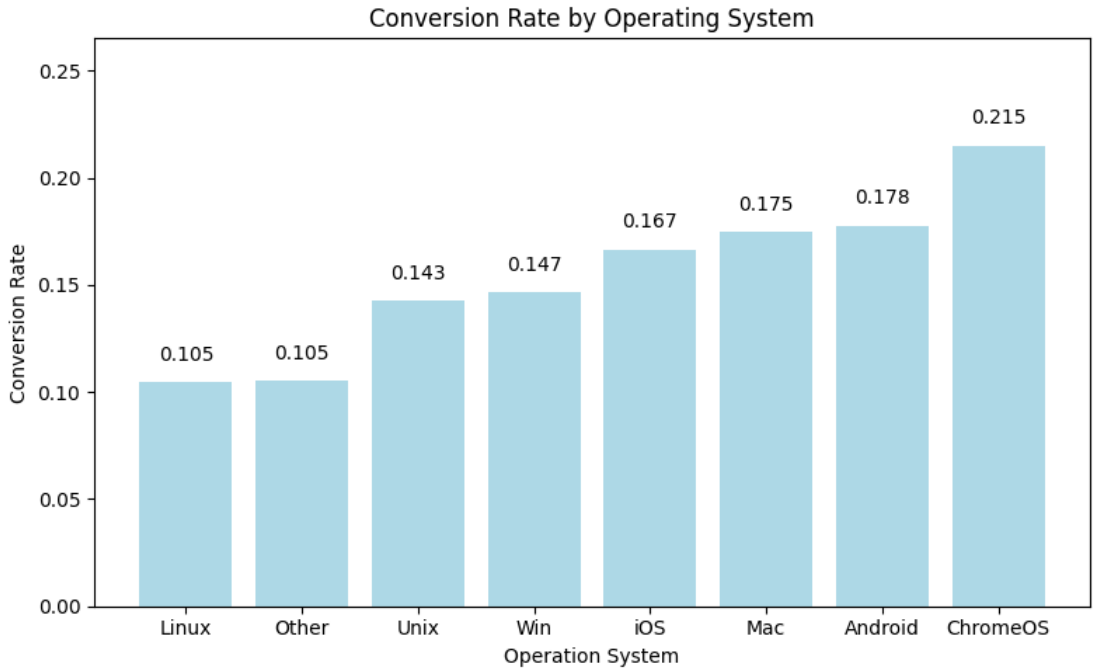
- Conversion rate by visitor type

We analyzed the conversion rate by visitor type to understand user engagement patterns. The data reveals that new visitors exhibit the highest conversion rate (0.25), followed by other visitor types (0.20), while returning visitors show the lowest conversion rate (0.14). This suggests that first-time users may have stronger purchase intent, whereas returning visitors may be more selective or information-seeking.



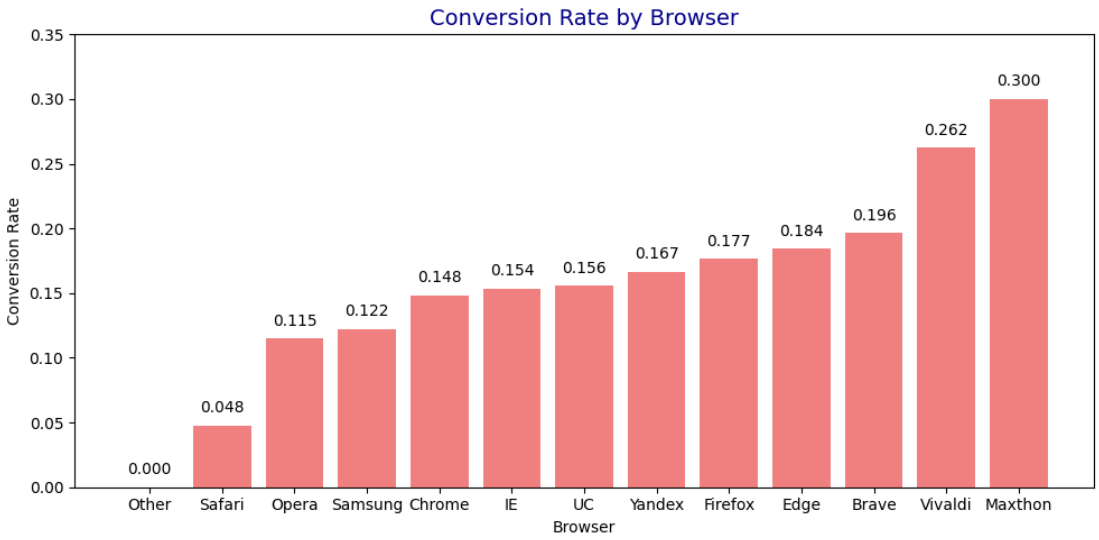
- Conversion rate by operation system

We further examined conversion performance by operating system. Users on ChromeOS demonstrated the highest conversion rate (0.215), followed by Android and Mac users. On the other hand, Linux and "Other" OS users showed the lowest rates (0.105). These results may reflect differences in platform usability, device type, or user demographics.



- Conversion rate by browser

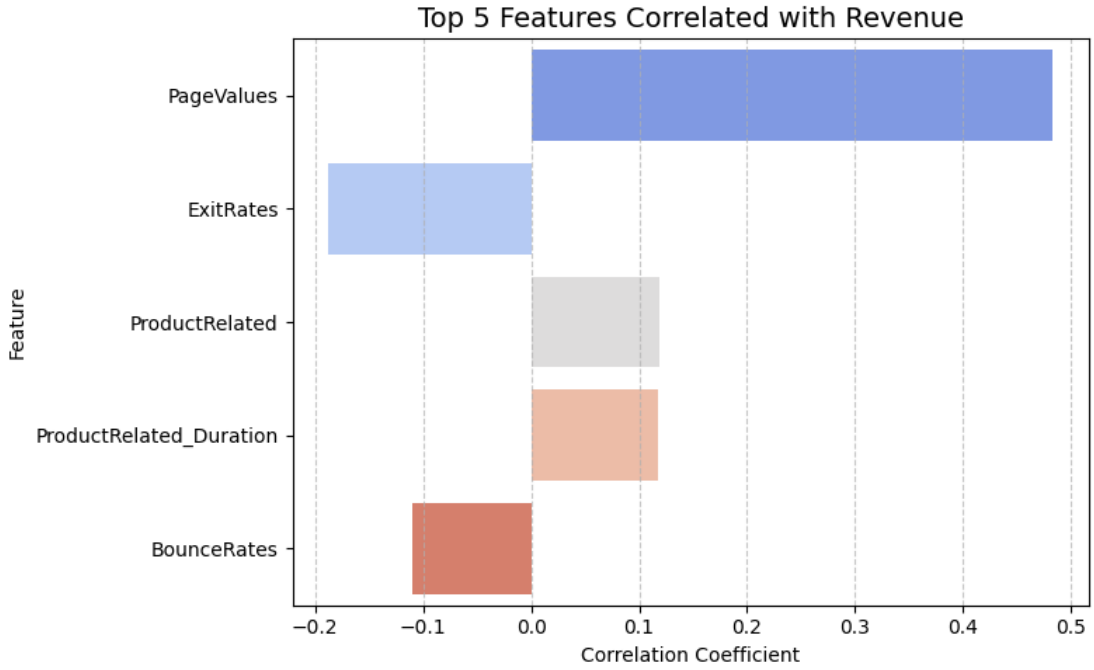
Analyzing browser-specific conversion rates shows that Maxthon users convert at the highest rate (0.300), followed by Vivaldi and Brave. In contrast, Safari and "Other" browsers had the lowest conversion rates. These findings may indicate browser performance differences or unique user behavior across platforms.



3. Correlation Analysis

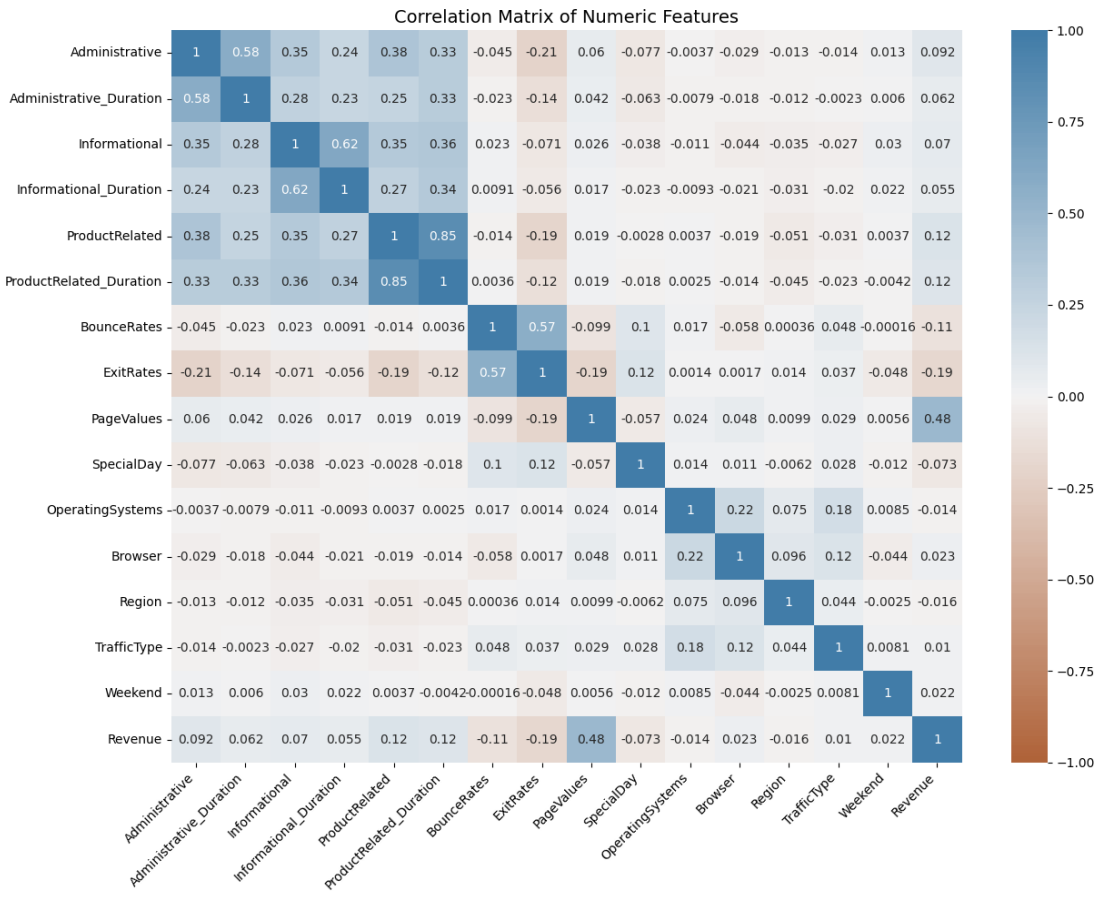
- Correlation matrix with revenue

To determine which features are most closely associated with revenue generation, we computed the correlation coefficients between all numeric variables and the revenue target. As shown in the bar chart, the top five most correlated features include PageValues (0.48), ExitRates (−0.19), ProductRelated (0.12), ProductRelated_Duration (0.12), and BounceRates (−0.11). Notably, PageValues has the strongest positive correlation with revenue, suggesting that users who engage with more product-related content are more likely to convert. Conversely, higher bounce or exit rates are associated with lower conversion probability.



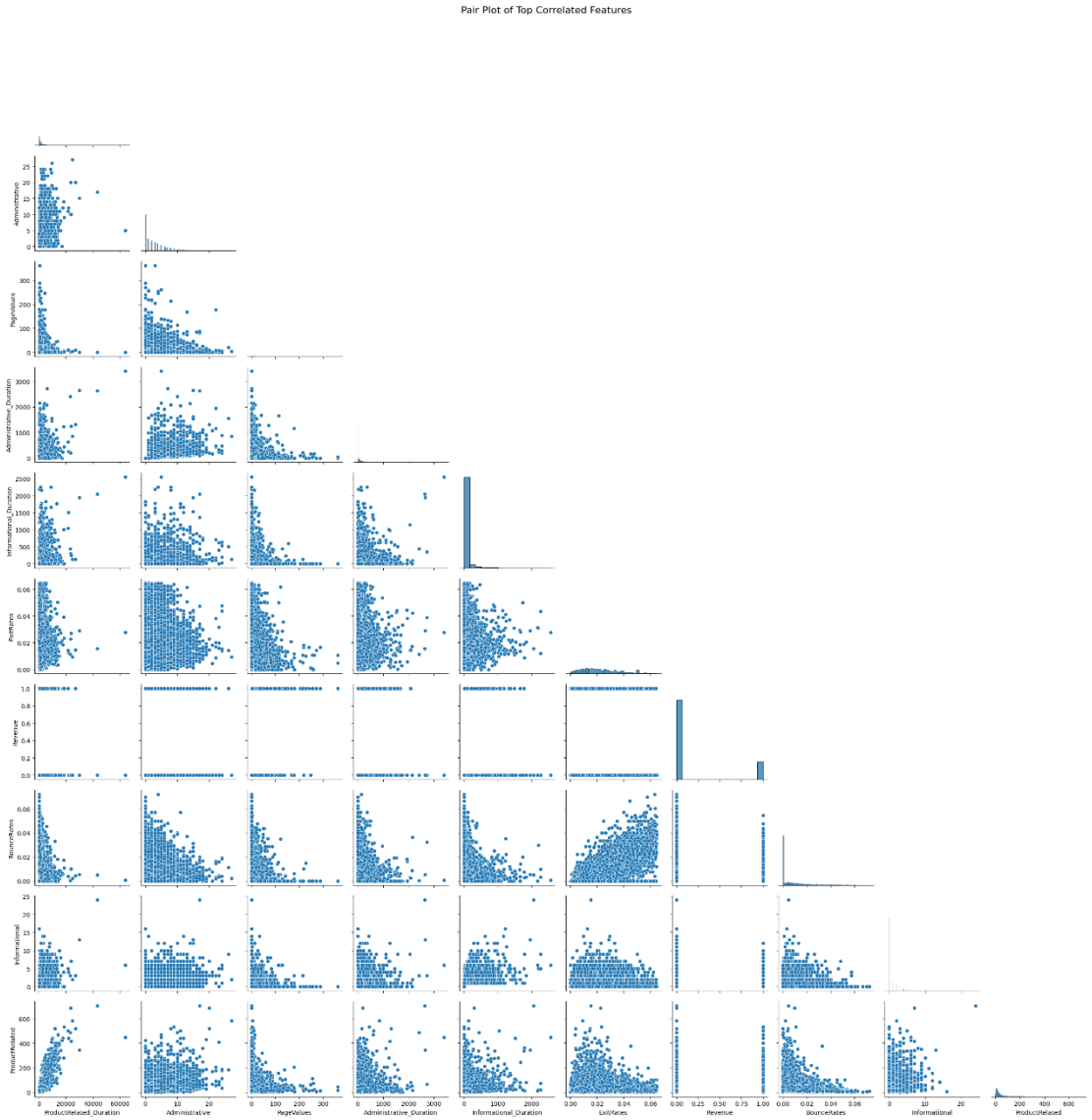
- Correlation matrix of numeric variables

We generated a full correlation matrix heatmap to analyze the relationships between all numeric variables. The matrix reveals several strong linear relationships—for example, ProductRelated and ProductRelated_Duration exhibit a strong positive correlation of 0.85, indicating that users who view more product-related pages also tend to spend more time on them. Similarly, Administrative and Administrative_Duration also show a significant correlation. This matrix helps identify redundant variables and potential multicollinearity, which are essential considerations in later model building.



- Top strongest/negative correlated variable pairs

We extracted the top 5 strongest positive and negative feature pairs based on absolute correlation values. The most positively correlated pair is again ProductRelated and ProductRelated_Duration (0.85), while the most negatively correlated pair is Administrative and ExitRates (−0.21). To further visualize the relationships, we constructed a pair plot displaying scatter plots and distributions among these top features. These visualizations confirm the linear associations and offer an intuitive understanding of feature interactions.



Business Insights and Conclusion

- Analytical Insights

The analysis revealed key behavioral traits influencing online purchasing decisions. New visitors exhibited significantly higher conversion rates, especially during months aligned with promotional periods. Users from specific regions and operating systems showed notable engagement, and session-level metrics such as PageValues and ExitRates strongly correlated with transaction likelihood. This underscores the importance of improving session quality and tailoring the experience for different visitor types to increase conversions.

- Strategy recommendations

To enhance conversion, businesses should prioritize engagement strategies for returning users, such as personalized promotions or loyalty rewards. Region-specific campaigns and OS-adaptive interface optimizations could also drive performance. Additionally, focusing on pages with high exit rates and enhancing content where PageValues are high may yield direct improvements in transaction outcomes.