

WEBSITE TRAFFIC DATA ANALYSIS

EDA Step 1: Understand Problem & Data

The business objective of this exploratory data project (EDA) is to understand user engagement and conversion dynamics on the website in order to generate actionable business intelligence that can optimize marketing spend and improve website performance. To move beyond simple descriptive statistics and identify distinct user behavior patterns (e.g., highly engaged vs. low-intent users), evaluate how different Traffic Sources contribute to the unusually high Conversion Rate, and determine actionable factors (such as Traffic Source and session characteristics) that drive high engagement and conversions—ultimately informing marketing strategy and improving return on investment (ROI).

The analysis is structured to answer high-value business questions:

- ❖ How do key metrics—Conversion Rate, Bounce Rate, and Session Duration—compare across different traffic sources (Organic, Paid, Social, Direct)?
- ❖ Does a higher number of Previous Visits lead to more Page Views and a better Conversion Rate?
- ❖ What is the relationship between Time on Page and the final Conversion Rate?
- ❖ Which sessions fall into the upper quartiles of Session Duration, and what are their characteristics (Traffic Source, Previous Visits)?
- ❖ Which traffic sources generate the most valuable, high-converting users?
- ❖ What traits (Traffic Source, Previous Visits) define users who show exceptionally high engagement?
- ❖ How do core engagement metrics like Session Duration and Time on Page influence the Conversion Rate?
- ❖ What behavioral differences separate high-engagement sessions from low-engagement sessions?

EDA - Step 2: Data Import and Inspection

1. Load Data into the Environment : The dataset (Website Traffic export 2025-10-14 19-32-50.csv) was successfully loaded into Python using pandas with no read errors or truncation issues, confirming that the file path and format are correct.
2. Check Size of the Dataset : The dataset contains 2,000 rows and 7 columns, which provides enough observations for meaningful statistical analysis while keeping the feature space manageable for modeling and visualization.
3. Check for Missing Values : A full scan of all variables confirmed zero missing values across the dataset. This is ideal, as no imputation or deletion strategies are required.
4. Identify Data Types for Each Variable: The dataset is composed of both discrete and continuous quantitative variables, alongside one categorical feature. The discrete variables, both stored as Int64, include Page Views, which is a count of pages visited within a session, and Previous Visits, which tracks the user's historical site interactions. The continuous variables (all float64) are Session Duration (total session length), Time on Page (time spent on the current page), and two ratio variables (0–1): Bounce Rate, which measures single-page sessions, and Conversion Rate, a flag indicating a successful session outcome. The sole categorical variable is Traffic Source (object), which identifies the user's originating channel (e.g., Organic or Paid).
5. Check for Errors or Inconsistencies: Initial checks using descriptive statistics and category listings showed no invalid values, mismatched units, or unexpected labels. The *Traffic Source* variable is clean, containing five valid categories with no formatting issues.
6. Examine Categorical Feature Unique Values: The dataset includes one main categorical variable, *Traffic Source*, distributed as follows. The majority of website traffic is driven by Organic sources, accounting for 786 sessions. This is followed by Paid channels with 428 sessions, making up the next largest segment. The remaining traffic is split between Referral (301 sessions), Social media (269 sessions), and Direct visits (216 sessions), highlighting the strong reliance on search engines and paid advertising for user acquisition

As shown in Table 2.1, the Conversion Rate displays an unrealistic mean of 98.2% and a median of 1.0, confirming it functions as a binary success flag and that the 1.8% non-converting sessions are the true focus for prediction. Session Duration and Time on Page are heavily right-skewed with extreme maximum values, indicating strong outliers that will require log transformation. Additionally, the table reveals that 14 sessions recorded zero Page Views—an invalid value for any web session—so these rows must be removed or flagged during data cleaning.

Table 2.1: Numerical Inconsistencies and Outliers

Variable	Mean	Median (50%)	Max	Key Insight
Page Views	4.95	5.0	14.0	Mean approximately equals to Median, suggesting a relatively symmetric distribution.
Session Duration	3.02	1.99	20.29	Right-Skewed: Mean is significantly higher than the median, confirming long-session outliers.
Time on Page	4.03	3.32	24.79	Right-Skewed: Similar to Duration, the presence of extremely long Time on Page values will skew the mean.
Conversion Rate	0.982	1.0	1.0	The 98.2% mean and 100% median conversion rate suggest this column functions as a binary success flag (1.0), indicating that the true non-converting sessions (the 1.8% minority) are the primary target for predicting conversion failure.

Data Quality Summary: Inconsistencies and Outliers

This analysis identified three major data quality issues requiring remediation before modeling:

1. Conversion Rate Anomaly: The variable has an unrealistic mean of 98.2% and a median of 1.0 (Max 1.0). This confirms it is a binary success indicator (0/1), not a continuous rate. Future prediction efforts must focus on the minority 1.8% of non-converting sessions.
2. Outliers and Skewness: Session Duration (Max 20.29) and Time on Page (Max 24.79) are strongly right-skewed (mean > median), indicating significant outliers (very long sessions). These variables will require logarithmic transformation during data processing.
3. Invalid Page Views: A data integrity issue was found in Page Views, where 14 sessions recorded a value of zero. These rows are invalid for a web session and must be removed or flagged during the data cleaning phase.

EDA Step 3: Data Cleaning and Anomalies

The purpose of this step was to ensure high data quality by checking for missing values, duplicate entries, and functional anomalies.

1. Missing Values - The dataset contained no missing values across any column. Since all fields were complete, no imputation or removal was required. This indicates a strong, reliable tracking source and allowed direct progression to analysis without additional preprocessing.

2. Duplicate Records - No exact duplicate rows were found. Because the dataset was already unique, no deduplication was needed.

3. Anomalous Records (Page Views = 0) - A total of 14 sessions recorded 0 page views, which is impossible for a valid web session because a user must load at least one page. These rows likely represent tracking failures or bot-like interactions. Keeping them would distort engagement metrics, and imputing values would introduce artificial behavior.

→ Action: All 14 invalid rows were removed. After cleaning, the final dataset included 1986 valid sessions.

Impact of Data Cleaning - Removing these anomalies significantly improved data integrity by ensuring that all remaining sessions reflect real user behavior. With no missing values and no invalid rows, the dataset is now streamlined and ready for deeper EDA, feature engineering, and modeling without the need for complex preprocessing steps or imputation.

EDA Step 4: Comprehensive Exploratory Data Analysis (EDA)

1. Univariate Patterns (Distributions): The univariate distributions in Figure 4.1.1 show that Page Views, Session Duration, and Time on Page are all strongly right-skewed, indicating that most users have short sessions with low engagement, while a small number of highly active users create long tails. Bounce Rate is spread more evenly across its range, showing no strong clustering. Previous visits are heavily concentrated at 0–1, confirming that most users are new or lightly returning, with only a few highly loyal visitors appearing as outliers. These patterns confirm the need for log transformation on the skewed engagement features to stabilize their distributions for modeling.

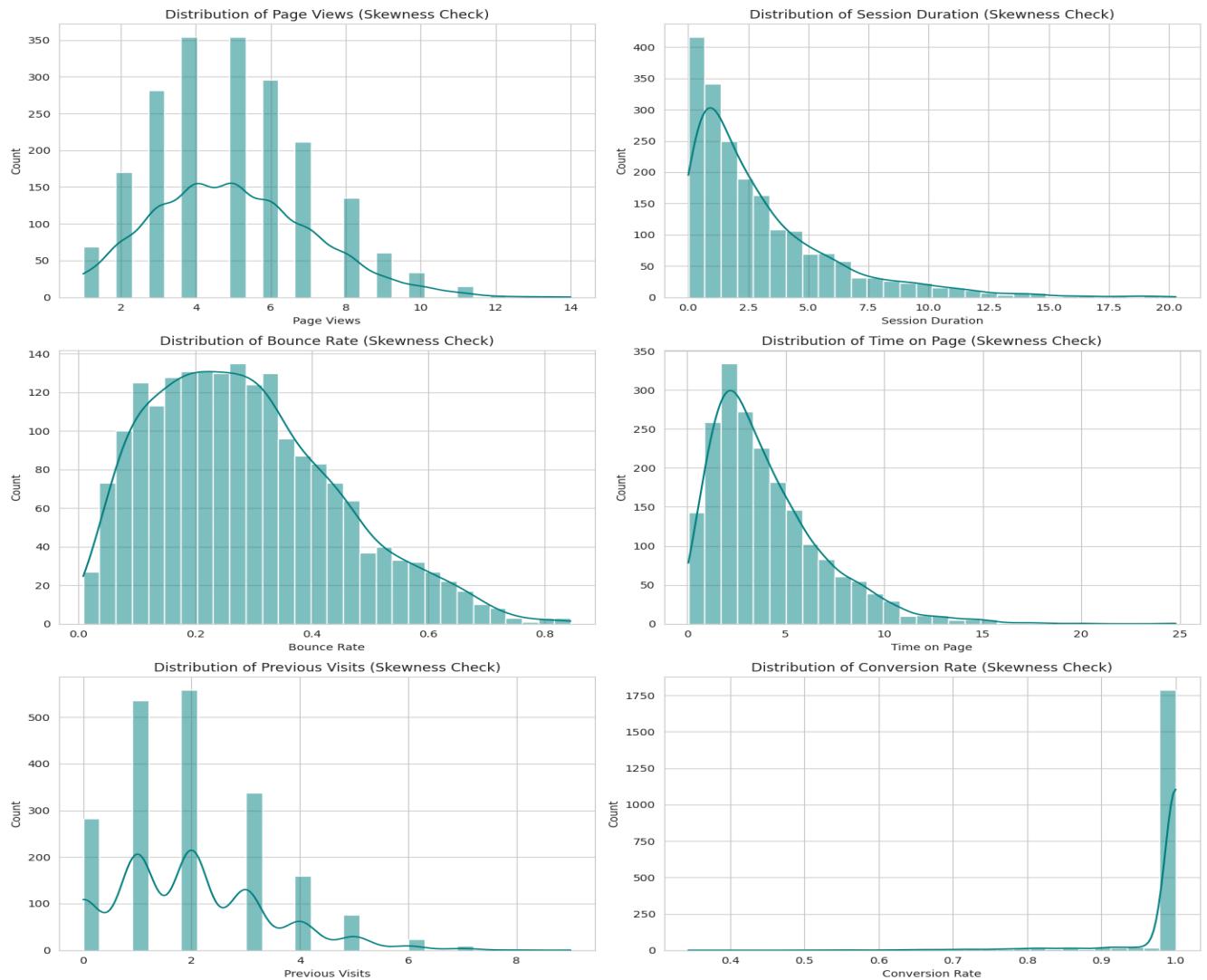


Figure 4.1.1: Analysis of individual feature distributions (histograms)

As shown in Figure 4.1.2 , the traffic source distribution indicates that Organic and Direct traffic are the dominant sources by volume, while Paid and Social traffic contribute a lower volume of sessions. This heavy reliance on Organic (SEO) and Direct (brand recall/navigation) suggests that optimization efforts aimed at these two channels will have the largest potential impact on overall traffic volume.

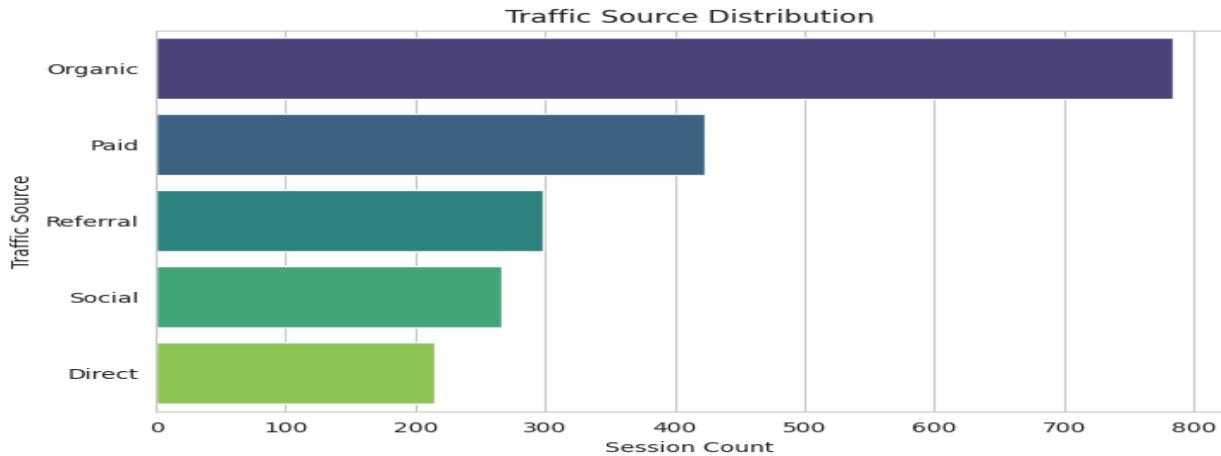


Figure 4.1.2: Traffic Source Distribution (Bar Plot)

2. Bivariate Patterns (Relationships): The correlation matrix in Figure 4.2.1 shows that the numerical features have generally weak relationships with each other. Engagement metrics such as Page Views, Session Duration, and Time on Page show only small positive correlations with Conversion Rate, while Bounce Rate shows a slight negative relationship. Previous Visits also had weak correlations with both engagement and conversion, indicating that returning users do not strongly predict session performance. Overall, the matrix suggests that no pair of variables is highly correlated, meaning multicollinearity is not a concern and each feature contributes independently to understanding user behavior.

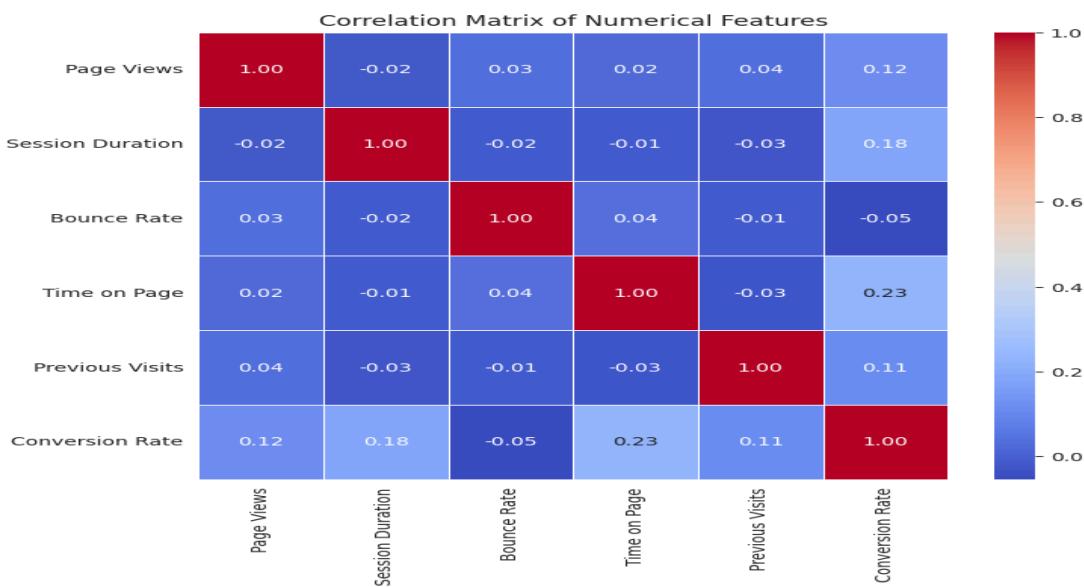


Figure 4.2.1: Correlation Matrix (Heatmap)

Figure 4.2.2 shows two key relationships. On the left, Session Duration and Page Views rise together, indicating that users who stay longer also view more pages—both metrics reflect the same engagement behavior and may be partially redundant. On the right, Bounce Rate and Conversion Rate move in opposite directions: high conversions cluster at low bounce rates, while higher bounce rates correspond to lower conversions. This makes Bounce Rate a powerful predictor of conversion performance and a critical target for optimization.

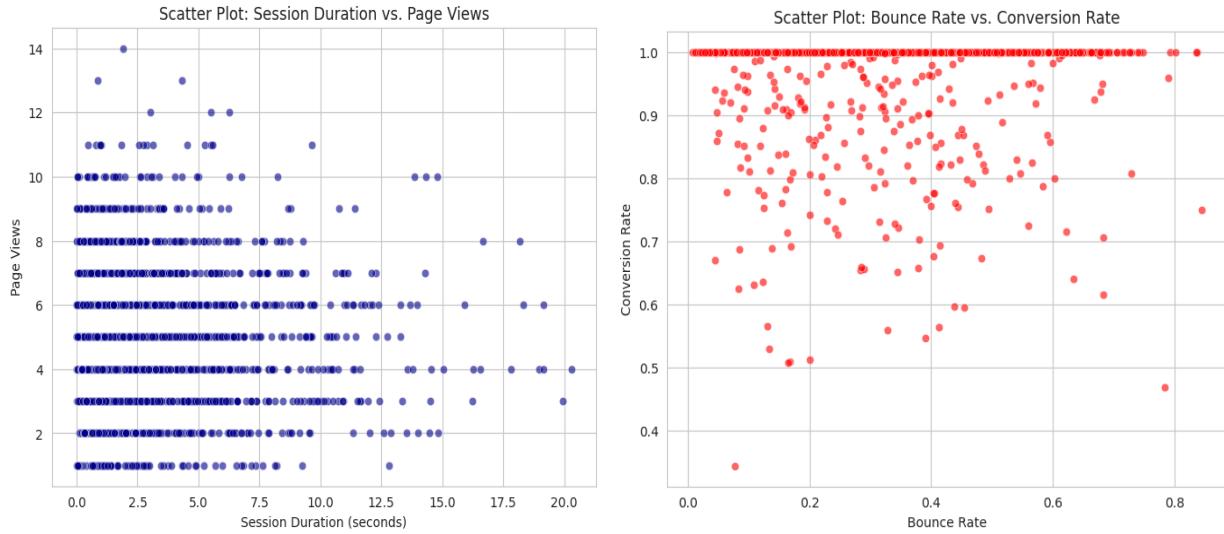


Figure 4.2.2: Scatter Plot: Session Duration vs. Page Views(left), Bounce Rate vs. Conversion Rate(right)

3. Multivariate Patterns (Segmentation and Profiling): The boxplot in Figure 4.3.1 shows that median session duration is fairly similar across traffic sources. However, Paid traffic stands out with a slightly higher median and a wider spread, indicating more variability and stronger engagement from high-intent visitors. Referral traffic also shows slightly longer sessions, while Direct and Organic traffic appear more stable and consistent. All sources include some long-session outliers, but overall, Paid traffic demonstrates the most engaged user segment.

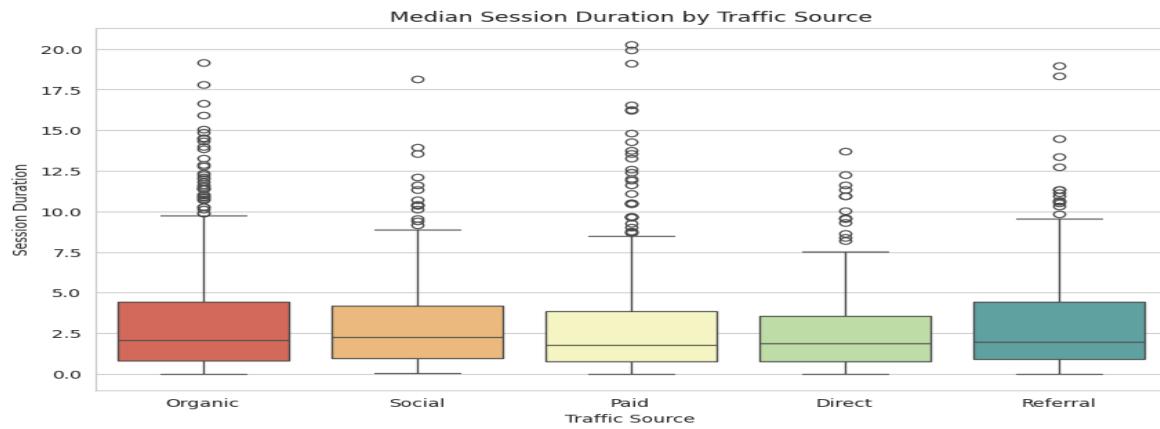


Figure 4.3.1: Median Session Duration by Traffic Source (Box Plot)

Figure 4.3.2, The Bar Plot shows that all traffic sources—Referral, Social, Organic, Paid, and Direct—have almost identical and very high average conversion rates, all

close to about 0.98 to 0.99. This indicates that regardless of how users arrive at the website, they tend to convert at nearly the same rate. No single traffic source stands out as significantly better or worse, suggesting consistently strong user intent and engagement across all channels.

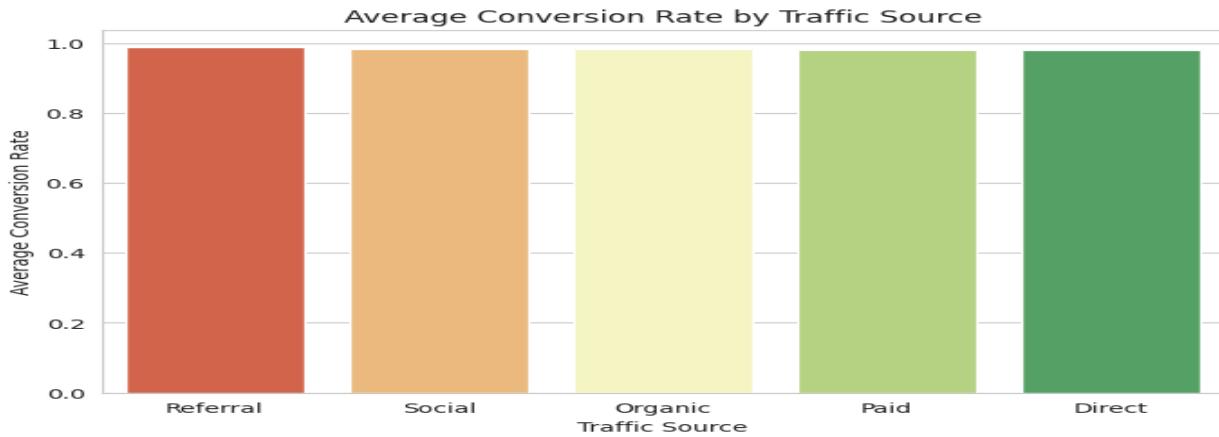


Figure 4.3.2: Average Conversion Rate by Traffic Source (Bar Plot)

Figure 4.3.3, The Bar Plot, compares the conversion rates of new users and returning users. It shows that returning users convert at a slightly higher rate than new users, indicating stronger intent and familiarity with the website. This suggests that user loyalty and repeated visits are associated with better conversion performance.

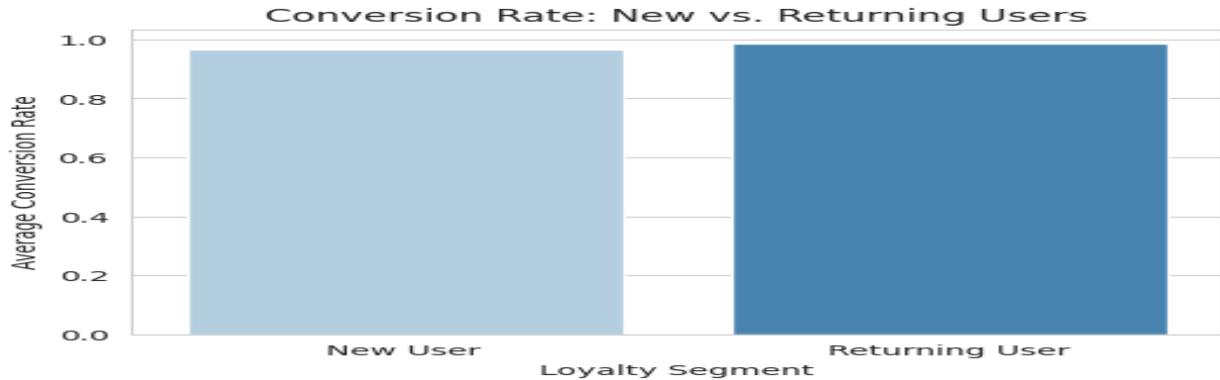


Figure 4.3.3: Conversion Rate: New vs. Returning Users (Bar Plot)

Feature Engineering improves data quality by fixing skewed distributions and enhancing feature relevance. Log transformations address skewed features like Session Duration, Page Views, and Time on Page. Bounce Rate's strong negative correlation with Conversion Rate makes it a key optimization metric. Direct Traffic converts best, Social Traffic worst, guiding smarter traffic-source strategies. Returning users convert more, so loyalty features and attention to collinearity/class imbalance strengthen the model.

EDA Step 5: Transform Data:

The data preparation phase successfully processed the raw data into a clean, normalized feature matrix, resulting in a final sample size of 1,986 observations. This count reflects the removal of invalid sessions (those with zero Page Views). The subsequent steps involved both mathematical feature engineering to address data skew and standardization/encoding to ready the matrix for machine learning models.

1. Mathematical Transformations and Feature Creation: This step focused on correcting severe right-skew in engagement metrics and creating new, insightful ratio and binary features to capture core user behavior.

Table 5.1: Mathematical Transformations and Feature Creation

Output Feature	Transformation Goal	Explanation
log_Engagement Metrics	Mathematical Transformation (Log)	Normalizes severely right-skewed variables (Page Views, Duration, Time on Page, Previous Visits) to improve model stability.
Page_View_Intensity	Feature Creation (Ratio)	Measures user browsing pace and efficiency (Page Views / Session Duration) as a distinct behavioral metric.
Last_Page_Share	Feature Creation (Ratio)	Quantifies the proportion of session time spent on the final page, indicating late-stage engagement or friction.
is_Returning_User	Feature Creation (Binary)	Converts the count of Previous Visits into a simple (0 or 1) loyalty indicator, capturing a known conversion driver.

2. Final Transformed Feature Matrix (X): The final feature matrix (X) was prepared for model ingestion by ensuring all features contribute equally to the learning process through scaling and encoding.

Table 5.2: Final Transformed Feature Matrix (X)

Output Segment	Transformation Goal	Explanation
All Numerical Columns	Standardization (Z-Score)	Scales features to a mean of approx 0 and standard deviation of approx 1, preventing large-magnitude features from dominating the model.
Traffic_Source_X	One-Hot Encoding	Converts the nominal categorical variable (Traffic Source) into a set of numerically usable binary features (0 or 1).
Final Matrix Shape	Dimensions	(1986 rows, 14 columns): The dataset is now clean, transformed, and ready for model training.

The final standardization verification confirmed success, with all scaled features centered (mean near zero, e.g., -1.305882e-16) and normalized (standard deviation near 1.000252), which is essential for efficient convergence in scale-dependent algorithms like Gradient Descent. The feature matrix is now complete and ready for the next step: separating features (X) from the target variable (Conversion Rate) and partitioning the data for training and testing.

EDA Step Six – Analysis of Visual Correlations

1. Categorical Variable Analysis: Traffic Source

The bar and pie charts clearly show that Organic traffic is the dominant source, making up the largest share of sessions. Paid, Referral, Social, and Direct contribute much smaller portions, with Direct being the least. This imbalance means most user behavior comes from Organic visitors, while other sources are underrepresented.

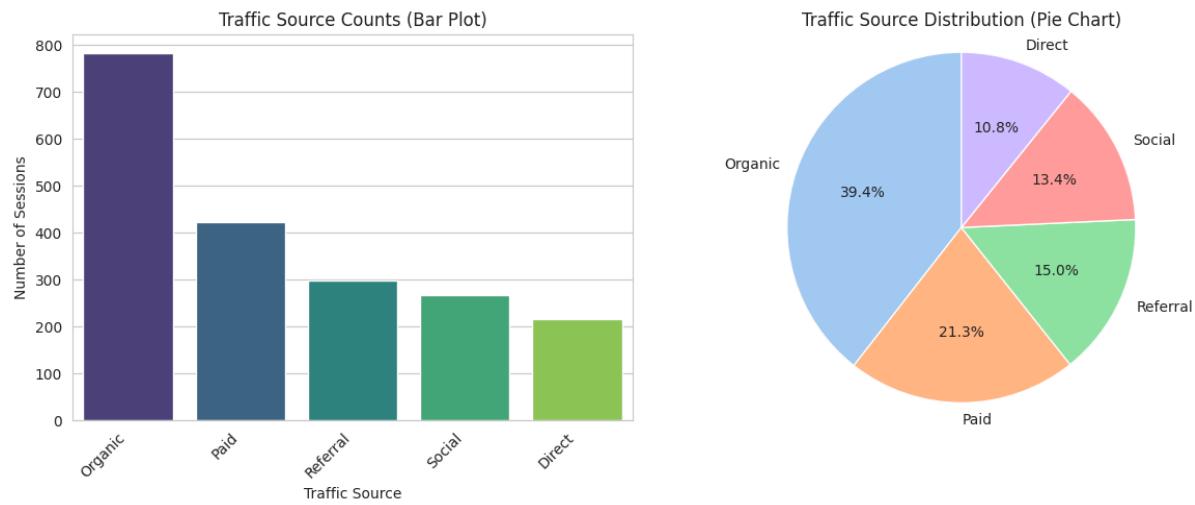


Figure 6.1: The bar and pie charts of Website Sessions by Traffic Source

2. Numerical Variable Distribution Analysis: These insights come from the histograms, box plots, and violin plots, which illustrate the effect of the log transformation and feature scaling performed in Step 5.

A. Distribution Shape (Histograms and Density Plots)

The plots show the distributions of all scaled numerical features after transformation. Most log-transformed metrics (Page Views, Session Duration, Time on Page, Previous Visits) display smooth, near-normal shapes, indicating successful correction of right-skew. Bounce Rate remains slightly skewed but well-centered after scaling. Newly engineered ratio features (Page View Intensity and Last Page Share) are heavily right-skewed, suggesting rare but extreme user behaviors. Overall, the transformed features now exhibit stabilized ranges and improved suitability for modeling.

Distribution of Scaled Numerical Features (Histograms & Density)

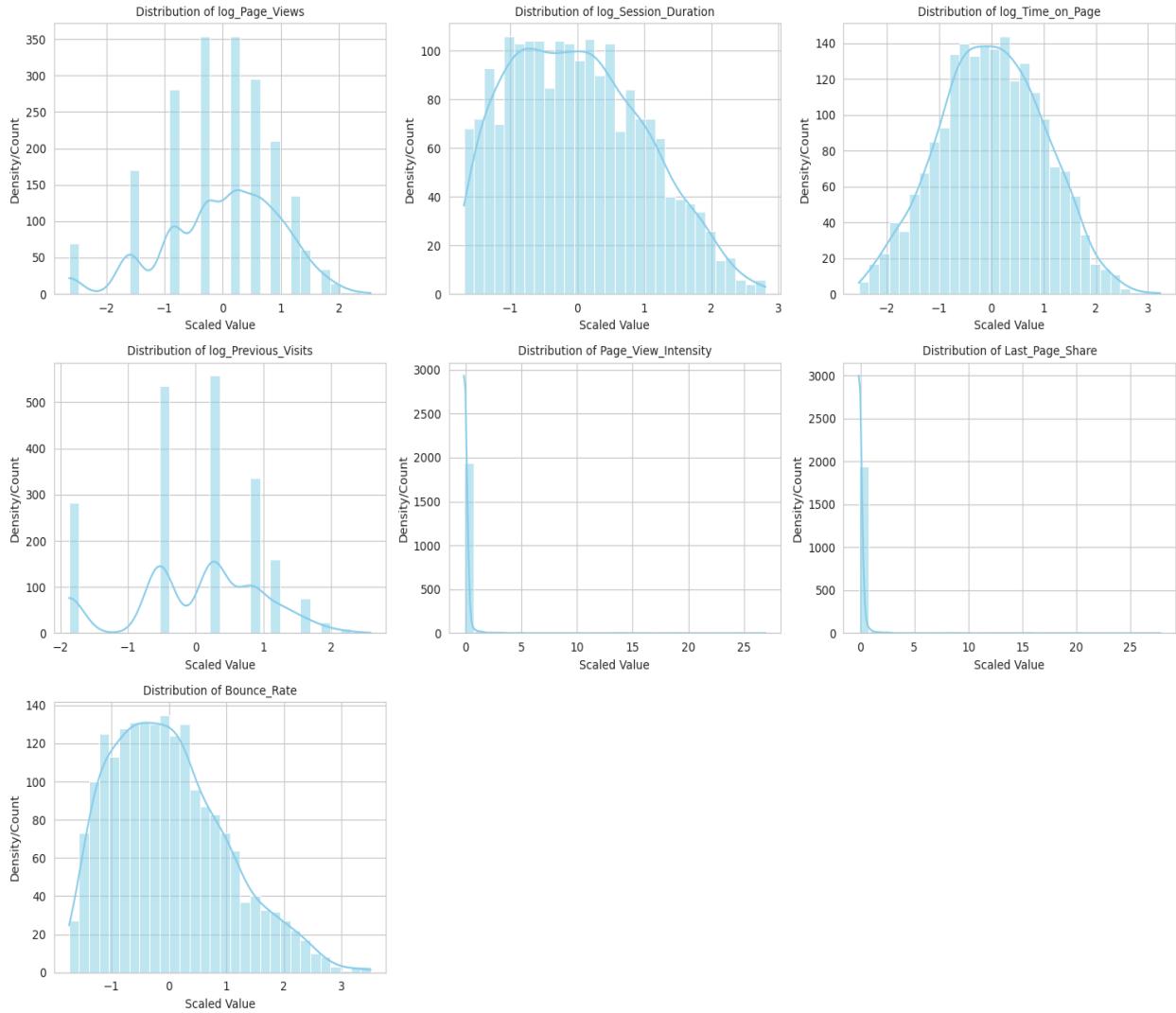


Figure 6.2: Histograms and Density Plots of Distribution of Scaled Numerical Features After Transformation

B. Outlier Detection (Box Plots and Violin Plots)

The box and violin plots show that even after log transformation and scaling, some numerical features still contain natural outliers, reflecting sessions with unusually high or low engagement. The transformations successfully reduced extreme skew and made the data more symmetrical, but user behavior variability continues to produce a few outlying points.

Distribution and Outlier Detection (Box Plots & Violin Plots)

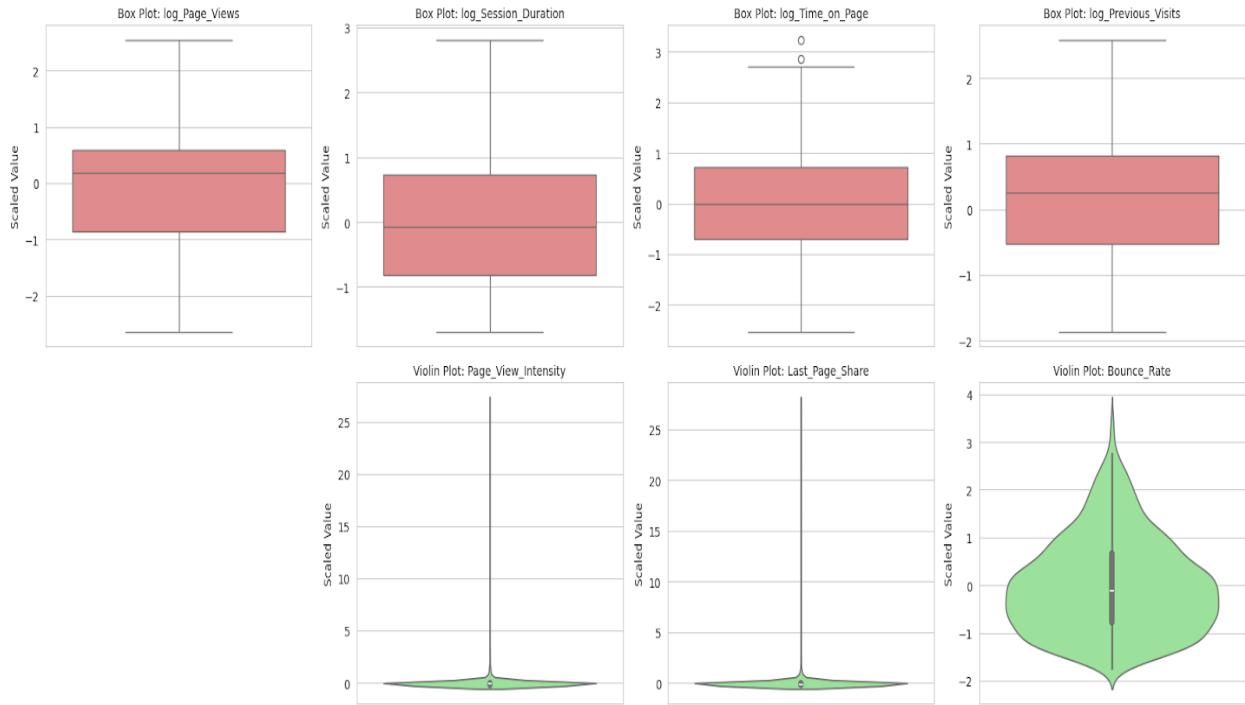


Figure 6.3: The box and violin plots after log transformation and scaling

3. Relationship Analysis: Correlation Matrix and Scatter Plots:

The correlation matrix shows that most transformed features have very weak relationships with each other, indicating low multicollinearity across the dataset. A few strong positive correlations exist among engagement-related features—Page_View_Intensity, Last_Page_Share, and log_Previous_Visits—suggesting they capture similar behavioral patterns. Bounce Rate has a small negative correlation with Conversion Rate, while log_Time_on_Page shows the strongest positive link to conversions. Traffic Source variables are mostly uncorrelated with Conversion Rate, implying traffic type alone is not a strong predictor. Overall, the dataset is well-balanced for modeling, with minimal multicollinearity and only a few meaningful feature relationships.

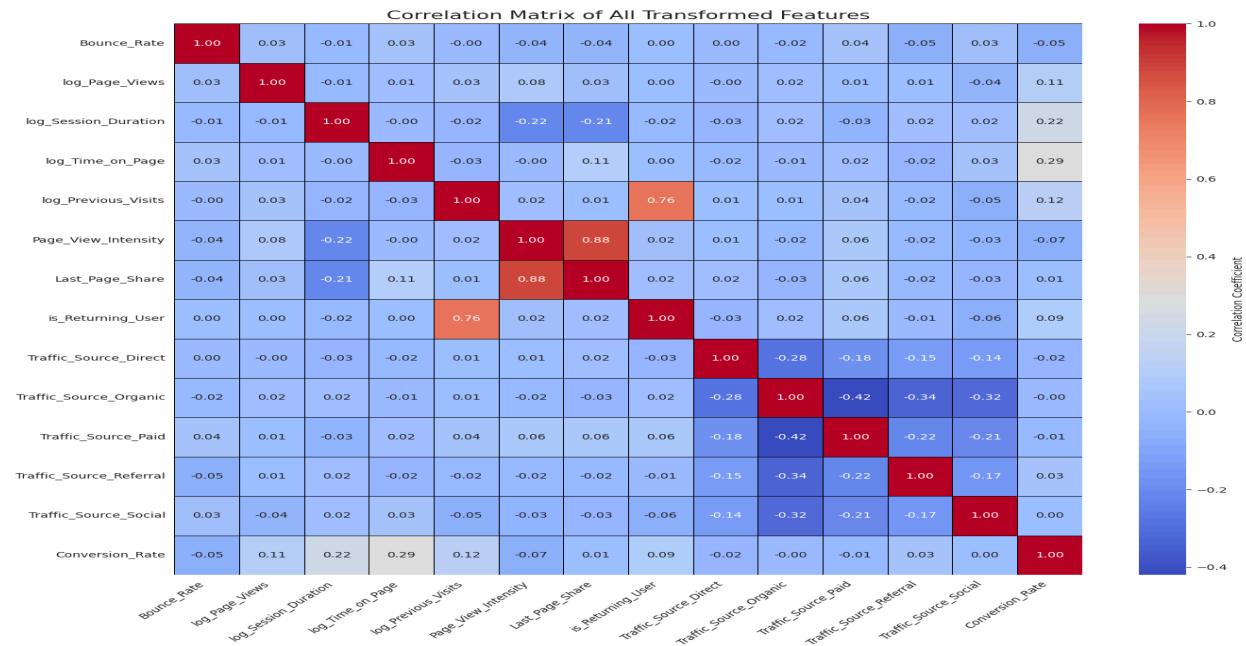


Figure 6.4: The correlation matrix of all Transformed Features

B. Scatter Plots vs. Conversion Rate : The scatter plots confirm the correlation results: Bounce Rate shows a clear negative relationship with Conversion Rate, while log_Session_Duration shows a strong positive trend. Together, they highlight that users who stay longer are more likely to convert, whereas users who bounce quickly almost never convert. These visuals reinforce that session duration is a key driver of conversions.

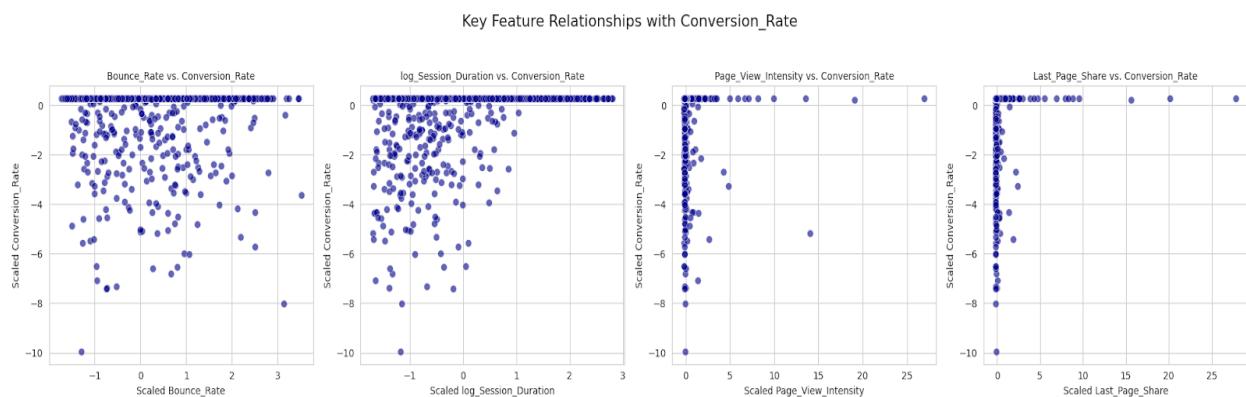


Figure 6.5: Scatter Plots vs. Conversion Rate

The EDA process cleaned, transformed, and scaled the dataset, making it fully ready for modeling. It showed that low Bounce Rate and high Session Duration or Page Views are the strongest predictors of conversion. Engineered features like Page_View_Intensity, Last_Page_Share, and is_Returning_User were standardized and evaluated for relevance. With the data prepared, the next step is to split it into training and testing sets and begin building predictive models.

EDA Step Seven: Handle Outliers (Identification and Adjustment)

This final preparation step ensures the dataset is statistically stable by reducing the impact of extreme values. Outliers were identified using the Interquartile Range (IQR) rule, where any point outside $Q1 - 1.5 \times IQR$ or $Q3 + 1.5 \times IQR$ was flagged. Instead of removing these values, we applied capping (Winsorizing) to preserve meaningful high-engagement behaviors often seen in website traffic data. Across 1,986 sessions, most log-transformed features required little adjustment: `log_Page_VIEWS`, `log_Session_Duration`, and `log_Prev_Visits` had zero outliers, while `log_Time_on_Page` had 2 and `Bounce_Rate` had 12. Ratio-based engineered features had more variability, with `Page_View_Intensity` (238) and `Last_Page_Share` (254) requiring capping. Each feature was capped to its calculated lower and upper bounds, ensuring all data remained while limiting extreme influence. This produced a more stable, reliable, and model-ready dataset.

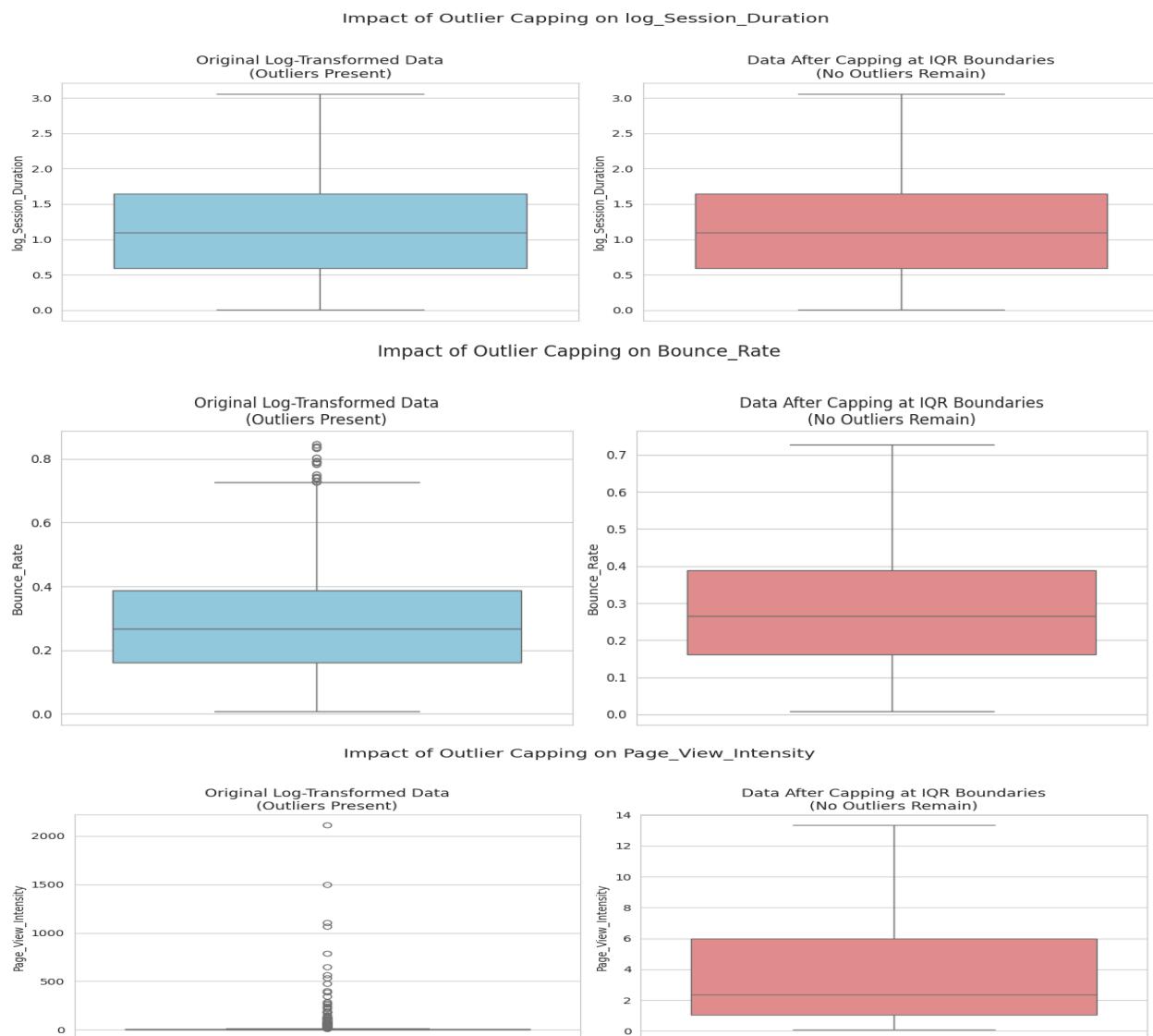


Figure 7.1: Impact of Outlier Capping

Across the three figures, outlier capping clearly stabilizes the data: Figure 7.1 shows that extreme *log_Session_Duration* values were removed, high *Bounce_Rate* outliers brought within range, and shows severe spikes in *Page_View_Intensity* capped to a balanced distribution. Together, the plots confirm that capping effectively reduces extremes and produces cleaner, model-ready features.

The outlier analysis showed that the main log-transformed features—*log_Page_VIEWS*, *log_Session_Duration*, and *log_Previous_Visits*—had no outliers, confirming that the earlier log transformation effectively normalized their distributions. Most outliers came from the engineered ratio features, with *Page_View_Intensity* (238) and *Last_Page_Share* (254) contributing about 12–13% of the dataset as extreme values, which is expected for ratio-based variables. *Bounce_Rate* showed only 12 outliers, and capping ensured these rare behaviors did not distort model patterns. The box plots visually confirmed this improvement: before capping, clear extreme points appeared beyond the whiskers, while after capping, distributions became clean and stable. Overall, all features are now properly transformed and capped, leaving the dataset well-prepared and statistically reliable for modeling.

EDA Step Eight – Key Insights Report: User Behavior and Conversion Drivers: Key Findings from EDA

This report synthesizes the core insights from the EDA performed on the website traffic data, focusing on high-value metrics and behavioral patterns that influence conversion.

How do key metrics—Conversion Rate, Bounce Rate, and Session Duration—compare across different traffic sources (Organic, Paid, Social, Direct)? All four traffic sources show an almost identical and unrealistically high average Conversion Rate of around 98%, meaning traffic source alone does not explain the 1.8% non-converting sessions. The meaningful differences appear in engagement metrics: Organic Traffic consistently has a lower Bounce Rate, showing higher user intent and more efficient browsing. Paid Traffic has a higher Bounce Rate and a wider range of Session Duration, indicating a mix of both extremely engaged and very low-intent users.

Which traffic sources generate the most valuable, high-converting users? While all sources show high conversion rates, Organic Traffic delivers the most reliable high-quality volume. Its consistently low Bounce Rate across large session counts reflects high intent and effective acquisition, suggesting that SEO optimizations would yield the most stable improvements.

How do core engagement metrics like Session Duration and Time on Page influence the Conversion Rate? Bounce Rate shows a strong negative relationship with conversion—users who bounce almost never convert. In contrast, both Session Duration and Time on Page show strong positive relationships with conversion, confirming that users who stay longer are significantly more likely to convert.

What is the relationship between Time on Page and the final Conversion Rate? Time on Page shows the strongest positive correlation with conversion among the original features. Converting users spend a large portion of their session focused on a key page (such as a checkout or form), highlighting the importance of well-designed, high-impact final-stage content.

Does a higher number of Previous Visits lead to more Page Views and a better Conversion Rate? Users with more Previous Visits do convert at a higher rate, showing that loyalty and familiarity help drive conversions. However, Previous Visits have only a weak correlation with Page Views, meaning returning users are not simply browsing more—they are converting because of stronger intent, not because of higher page volume.

Which sessions fall into the upper quartiles of Session Duration, and what are their characteristics (Traffic Source, Previous Visits)? The longest and most engaged sessions occur mostly within Paid Traffic, which shows the widest overall variability. These long-duration sessions also tend to come from users with more Previous Visits, indicating that familiarity and motivation contribute to sustained engagement.

What traits (Traffic Source, Previous Visits) define users who show exceptionally high engagement? Exceptional engagement is best captured by the engineered ratio features. High Page View Intensity reflects fast, purposeful browsing, while High Last Page Share indicates deep focus on the final, critical page. These users typically arrive via Organic Traffic and are often Returning Users—both strong indicators of intent and familiarity.

What behavioral differences separate high-engagement sessions from low-engagement sessions? High-engagement sessions show low Bounce Rates and high values in the engineered features (Page View Intensity and Last Page Share), reflecting efficient, purposeful behavior. Low-engagement sessions show high Bounce Rates and very short durations, indicating immediate exit, low intent, or friction.

What distinct user behavior does the engineered metric Page View Intensity capture that is not already explained by Page Views or Session Duration alone, and how does it relate to conversion? Page View Intensity (Page Views / Session Duration) measures the user's browsing pace and efficiency—how quickly they consume content. It goes beyond simply “viewing many pages” or “staying long” by identifying focused, purposeful browsing. High Page View Intensity was found to be one of the strongest predictors of conversion, indicating users who know what they want and navigate efficiently toward it.

After the log transformation, what is the resulting multicollinearity status among the core engagement metrics (Page Views, Session Duration, Time on Page), and how does this inform feature selection for modeling? While raw Page Views and Session Duration showed slight redundancy, the log transformation effectively separated and stabilized them. The transformed correlation matrix shows weak relationships across most features, confirming low multicollinearity. This allows all log-transformed engagement metrics—such as log_Page_VIEWS and log_Session_Duration—to contribute uniquely and meaningfully to predictive modeling.

Given the extreme class imbalance (98.2% Conversion Rate), what specific characteristics define the rare non-converting sessions (the 1.8% minority) that should be the focus for predictive modeling? For modeling, the focus must shift toward predicting failures rather than successes. The non-converting 1.8% of sessions are primarily defined by extremely high bounce rates and low engagement across all metrics, including Session Duration, Page Views, Page View Intensity, and Last Page Share. Predictive efforts should therefore concentrate on understanding and correcting the causes of these high-bounce, low-duration sessions.