# Website Traffic Data Analysis

Meenakshi Rajpurohit

# Business Objective & Key Questions

**Objective:** Understand user engagement and conversion dynamics to optimize marketing spend and improve website performance

**Key Questions to Answer:**

- Which traffic sources generate the most valuable, high-converting users?
- How do engagement metrics influence Conversion Rate?
- What traits define high-engagement vs low-engagement sessions?

# Dataset Overview

**Size:** 2,000 rows × 7 columns (Final: 1,986 after cleaning)

**Features:**

- **Continuous:** Session Duration, Time on Page, Bounce Rate, Conversion Rate
- **Discrete:** Page Views, Previous Visits
- **Categorical:** Traffic Source (Organic, Paid, Social, Referral, Direct)

**No missing values ✓**
**Data Quality Issue:** 14 sessions with 0 Page Views removed

# Data Quality & Cleaning

**Issues Identified:**

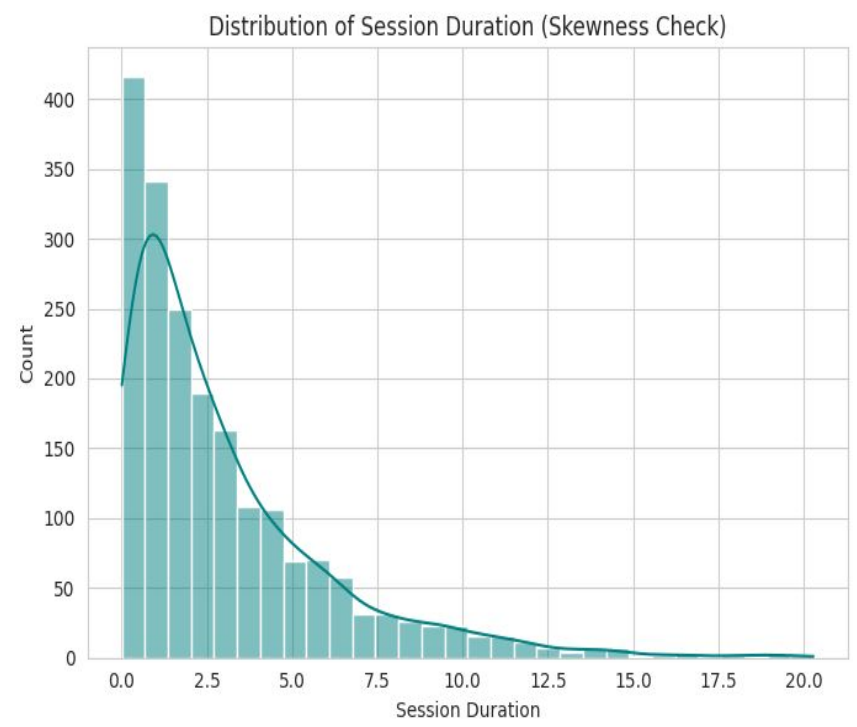
1. **Conversion Rate Anomaly:** 98.2% mean, 100% median → Binary flag (0/1)
2. **Extreme Outliers:** Session Duration (max: 20.29), Time on Page (max: 24.79)
3. **Invalid Records:** 14 sessions with 0 Page Views removed

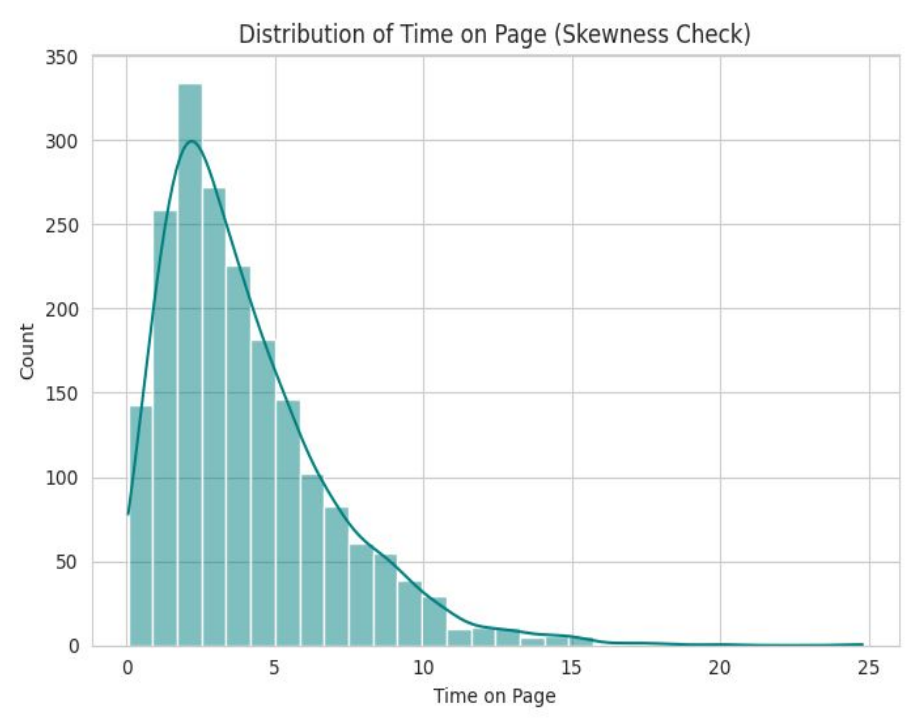
**Actions Taken:**

- Removed 14 invalid sessions
- Applied log transformation to skewed features
- Capped outliers using IQR method

| *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* | .982 | 1.0 | 1.0 | With a 98.2% mean and 100% median, this acts as a binary success flag. The key focus is the 1.8% minority of non-converting sessions |

# Issues Identified (Extreme Outliers):



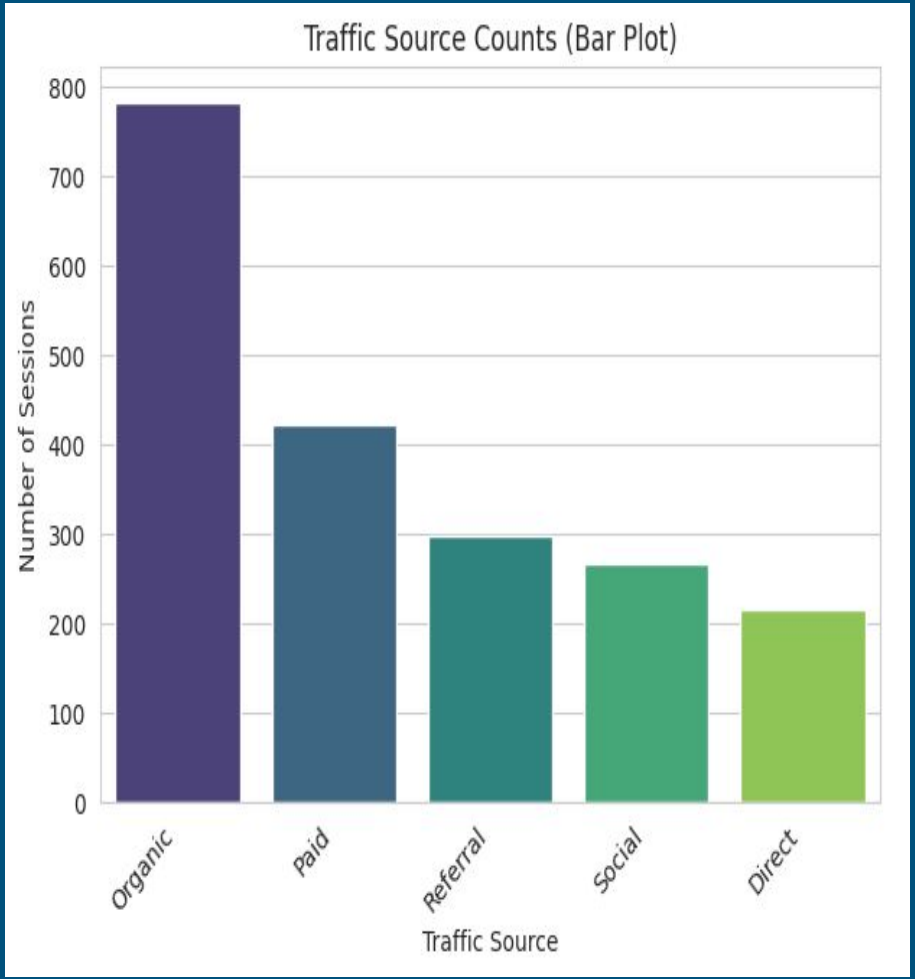



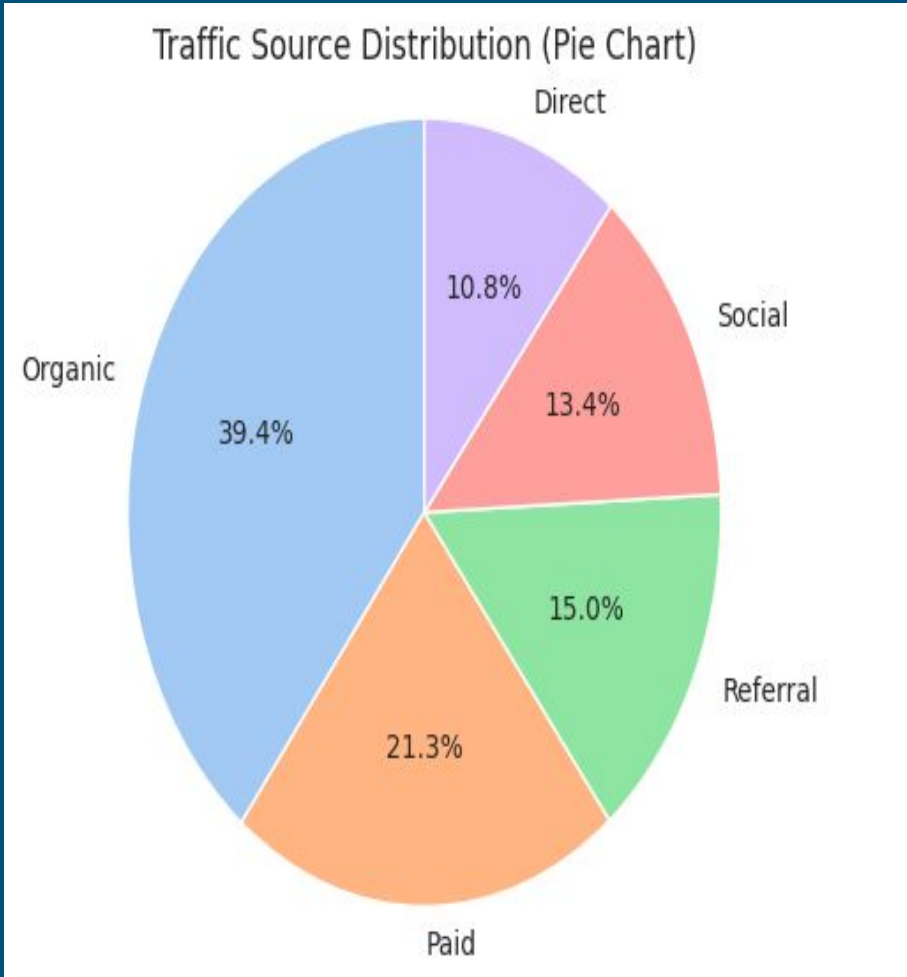
# Traffic Source Distribution

**Key Insights:**

- **Organic:** 786 sessions (39%) - Dominant source
- **Paid:** 428 sessions (21%)
- **Social:** 269 sessions (13%)
- **Referral:** 301 sessions (15%)
- **Direct:** 216 sessions (11%)

**Insight:** Heavy reliance on SEO and organic traffic

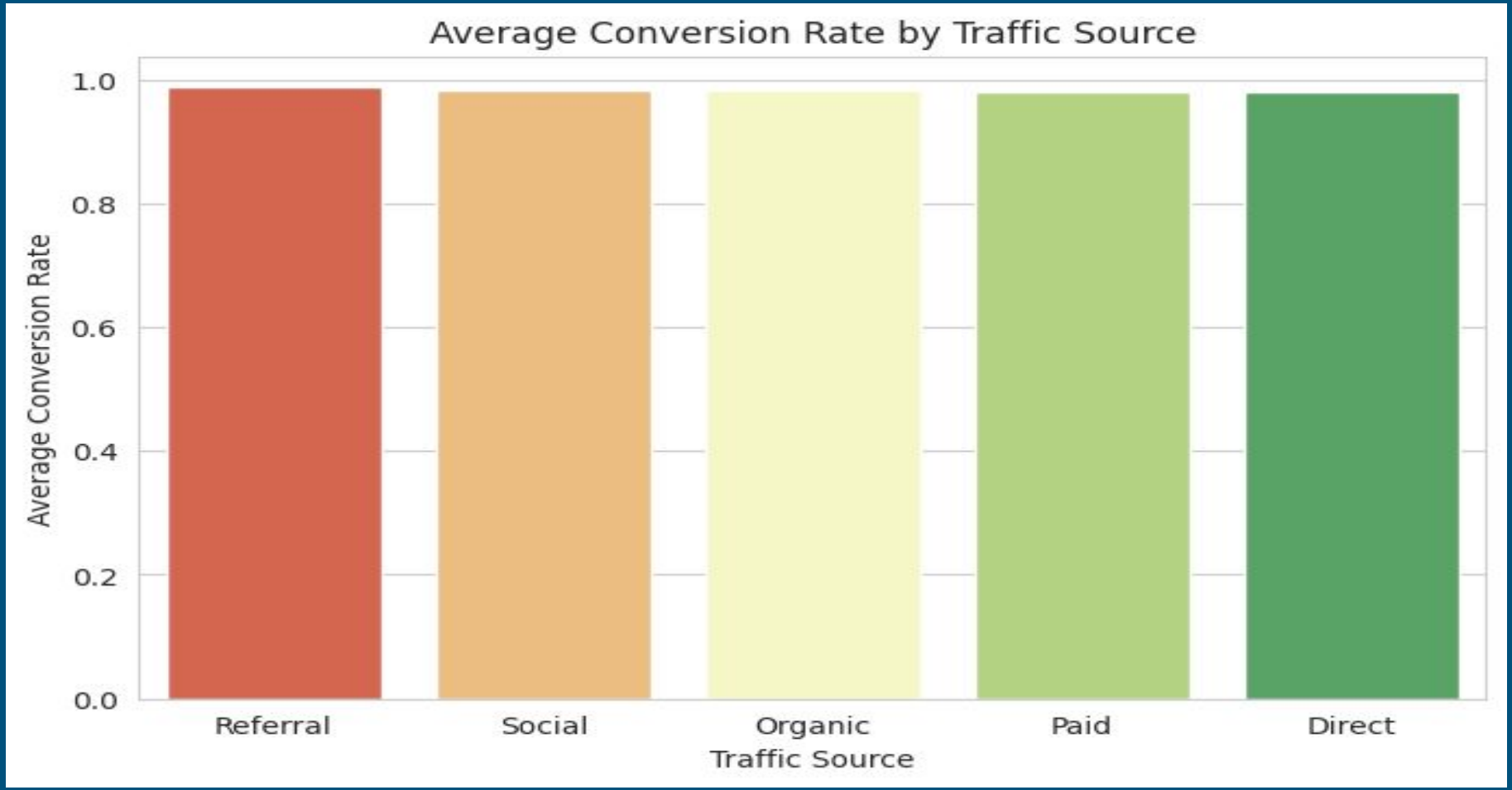

Traffic Source Counts (Bar Plot)
Number of Sessions
800
700
600
500
400
300
200
100
0
Organic
Paid
Referral
Social
Direct
Traffic Source


Traffic Source Distribution (Pie Chart)
Direct
10.8%
Social
13.4%
Organic
39.4%
15.0%
Referral
21.3%
Paid

# Conversion Rate Analysis by Traffic Source

**Key Finding:** All traffic sources show nearly identical high conversion rates (98-99%)

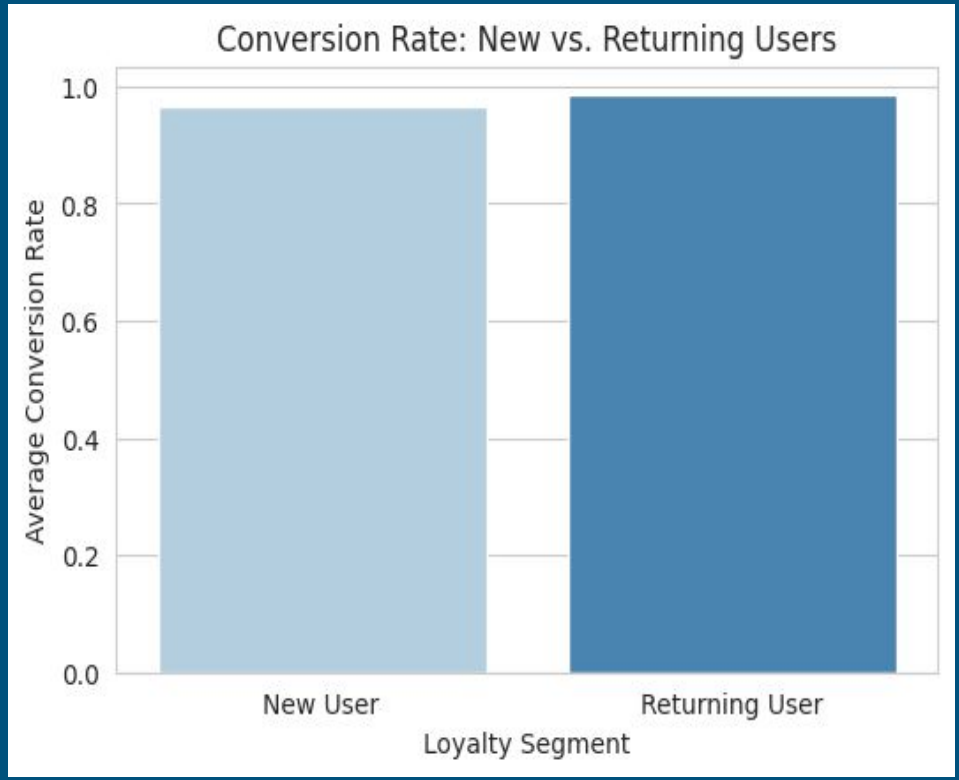
- Traffic source alone does NOT explain the 1.8% non-converting sessions
- Need to look at **engagement metrics** instead


Average Conversion Rate by Traffic Source
1.0
0.8
0.6
0.4
0.2
0.0
Average Conversion Rate
Referral
Social
Organic
Paid
Direct
Traffic Source

# Returning Users Impact

**Finding:** Returning users convert at slightly higher rates than new users

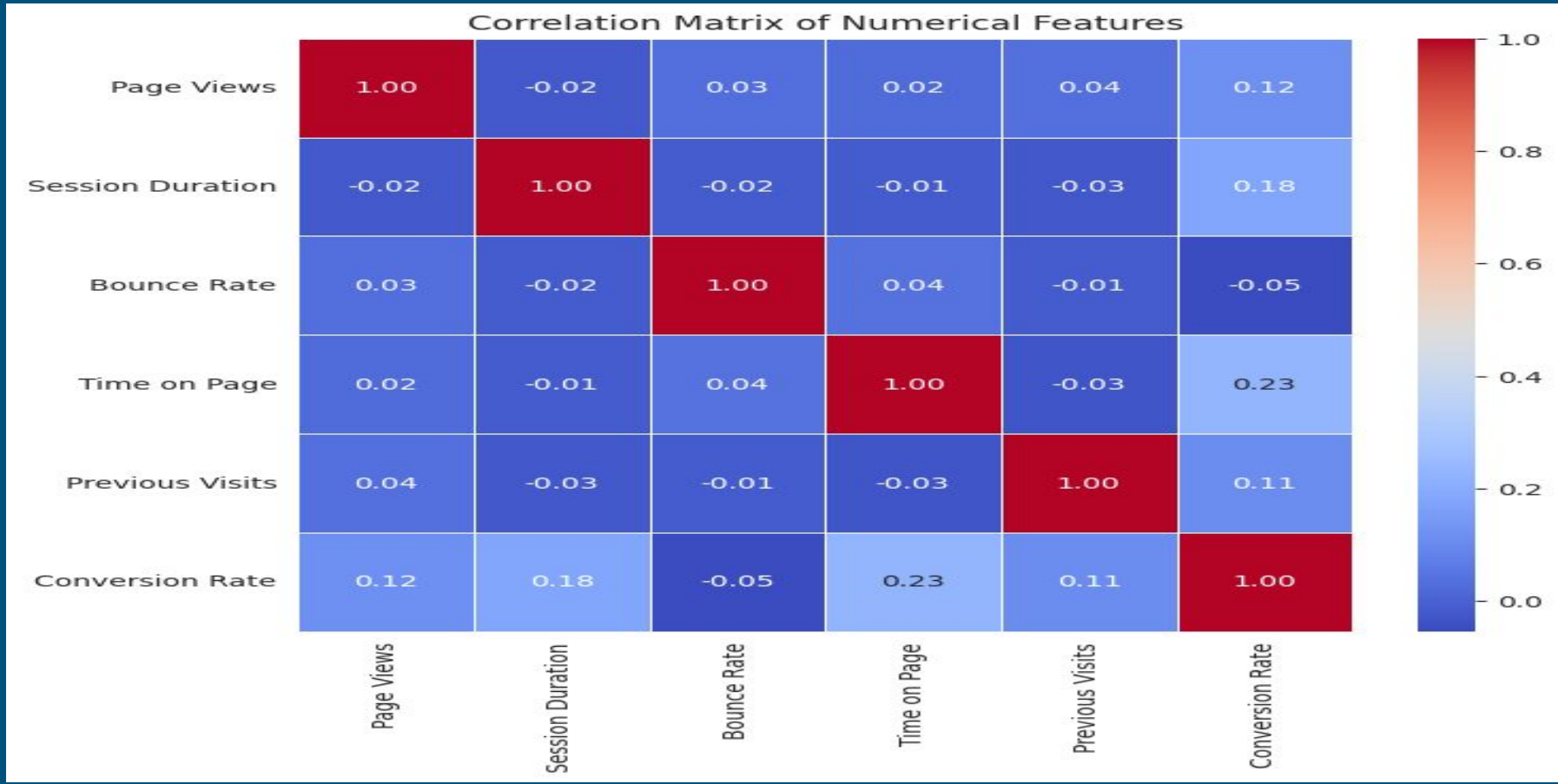
- Indicates stronger intent and familiarity with website
- **Recommendation:** Focus on retention strategies and loyalty programs

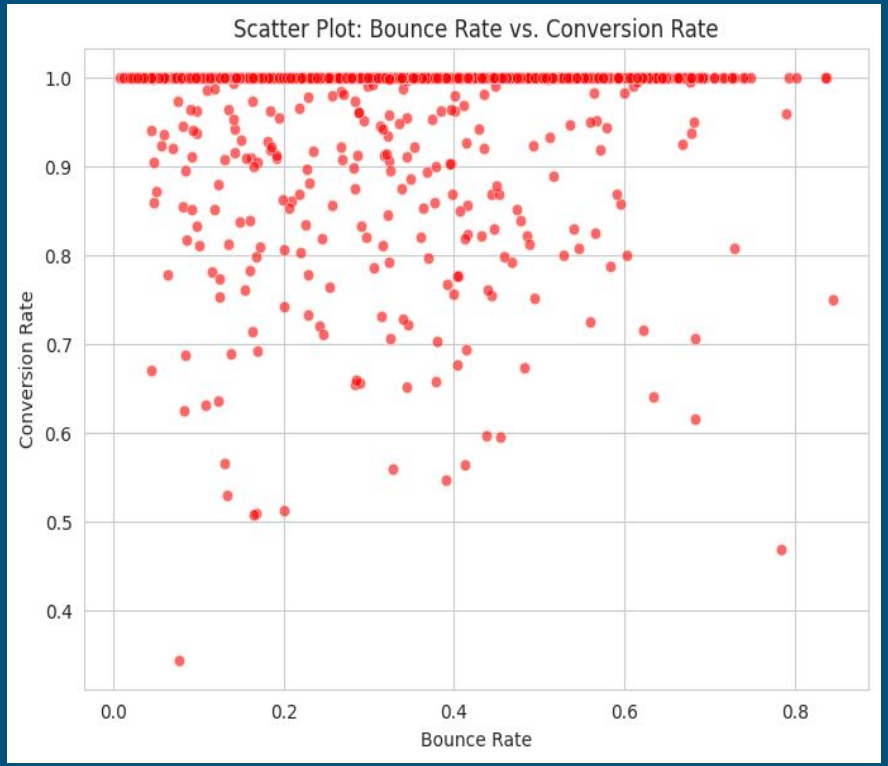


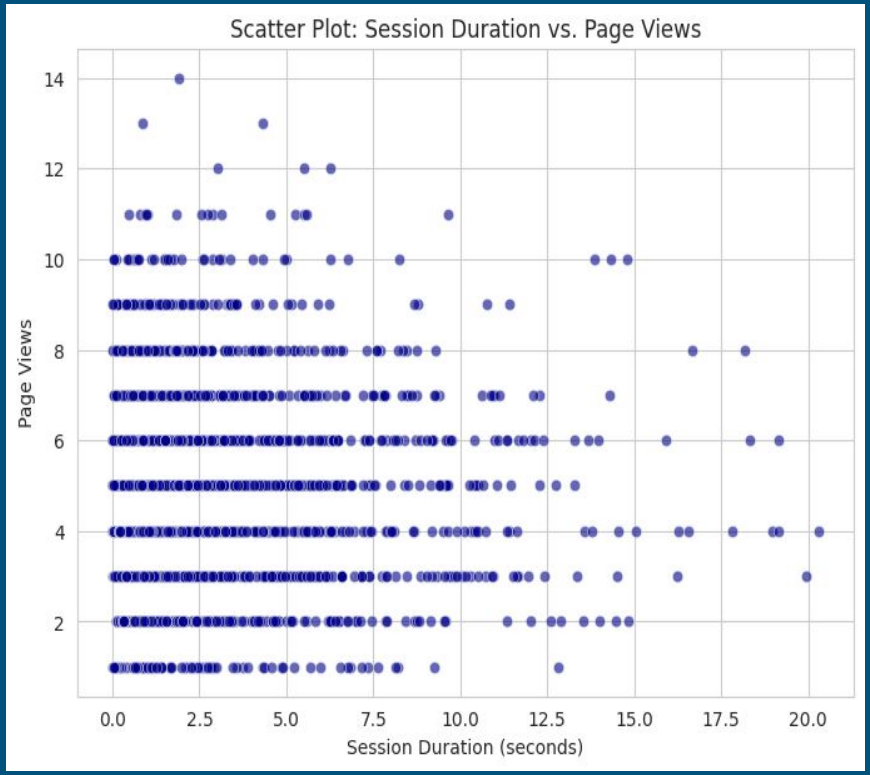
# Engagement Metrics - Key Correlations

**Critical Findings:**

1. **Bounce Rate ↔ Conversion Rate:** Strong negative correlation
    - High bounce = Almost never convert
2. **Session Duration ↔ Conversion Rate:** Strong positive correlation
    - Longer sessions = Higher conversion likelihood
3. **Time on Page ↔ Conversion Rate:** Strongest positive correlation
    - Users spending more time convert more


Correlation Matrix of Numerical Features
Page Views
Session Duration
Bounce Rate
Time on Page
Previous Visits
Conversion Rate
1.00 -0.02 0.03 0.02 0.04 0.12
-0.02 1.00 -0.02 -0.01 -0.03 0.18
0.03 -0.02 1.00 0.04 -0.01 -0.05
0.02 -0.01 0.04 1.00 -0.03 0.23
0.04 -0.03 -0.01 -0.03 1.00 0.11
0.12 0.18 -0.05 0.23 0.11 1.00
Page Views
Session Duration
Bounce Rate
Time on Page
Previous Visits
Conversion Rate
1.0
0.8
0.6
0.4
0.2
0.0


Scatter Plot: Bounce Rate vs. Conversion Rate
Conversion Rate
Bounce Rate
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8

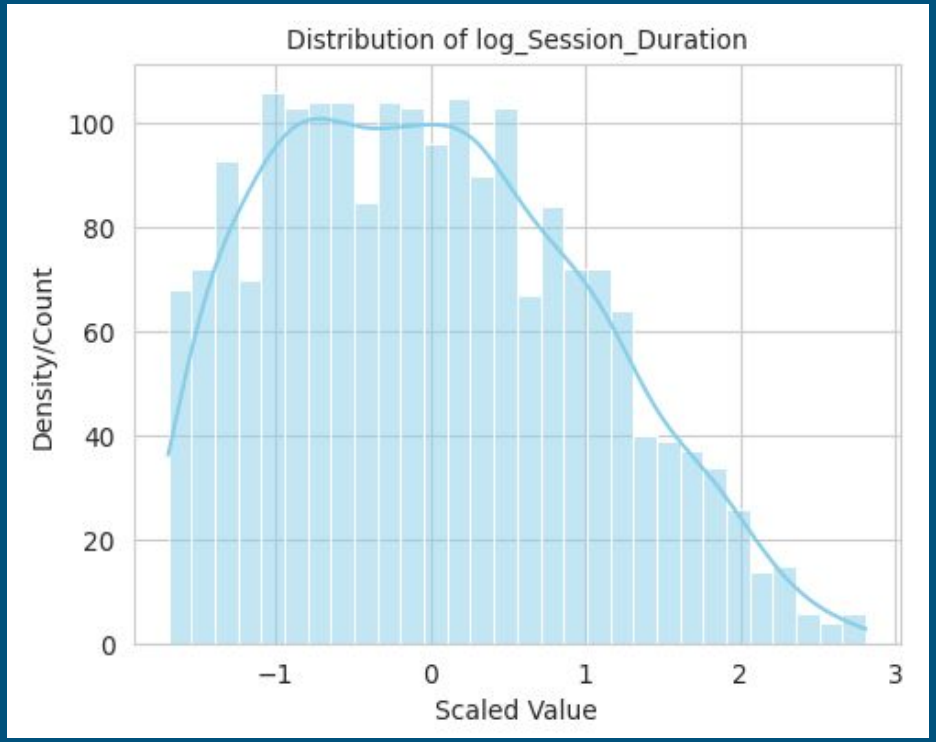

Scatter Plot: Session Duration vs. Page Views
Page Views
Session Duration (seconds)
14
12
10
8
6
4
2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

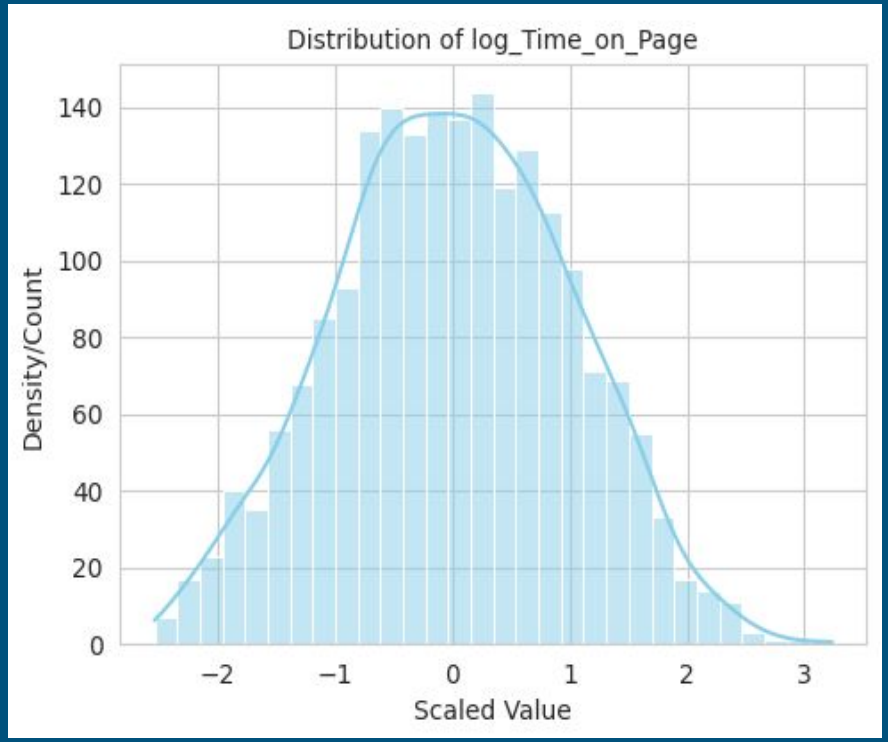
# Feature Engineering - New Insights

**Created 4 New Features:**

1. **Page View Intensity** = Page Views / Session Duration
   - Measures browsing efficiency and purposefulness
2. **Last Page Share** = Time on Page / Session Duration
   - Quantifies focus on final/critical page
3. **is_Returning_User** (Binary)
   - Returning users convert at slightly higher rates
4. **Log Transformations** on skewed features

# Distribution of Scaled Numerical Features

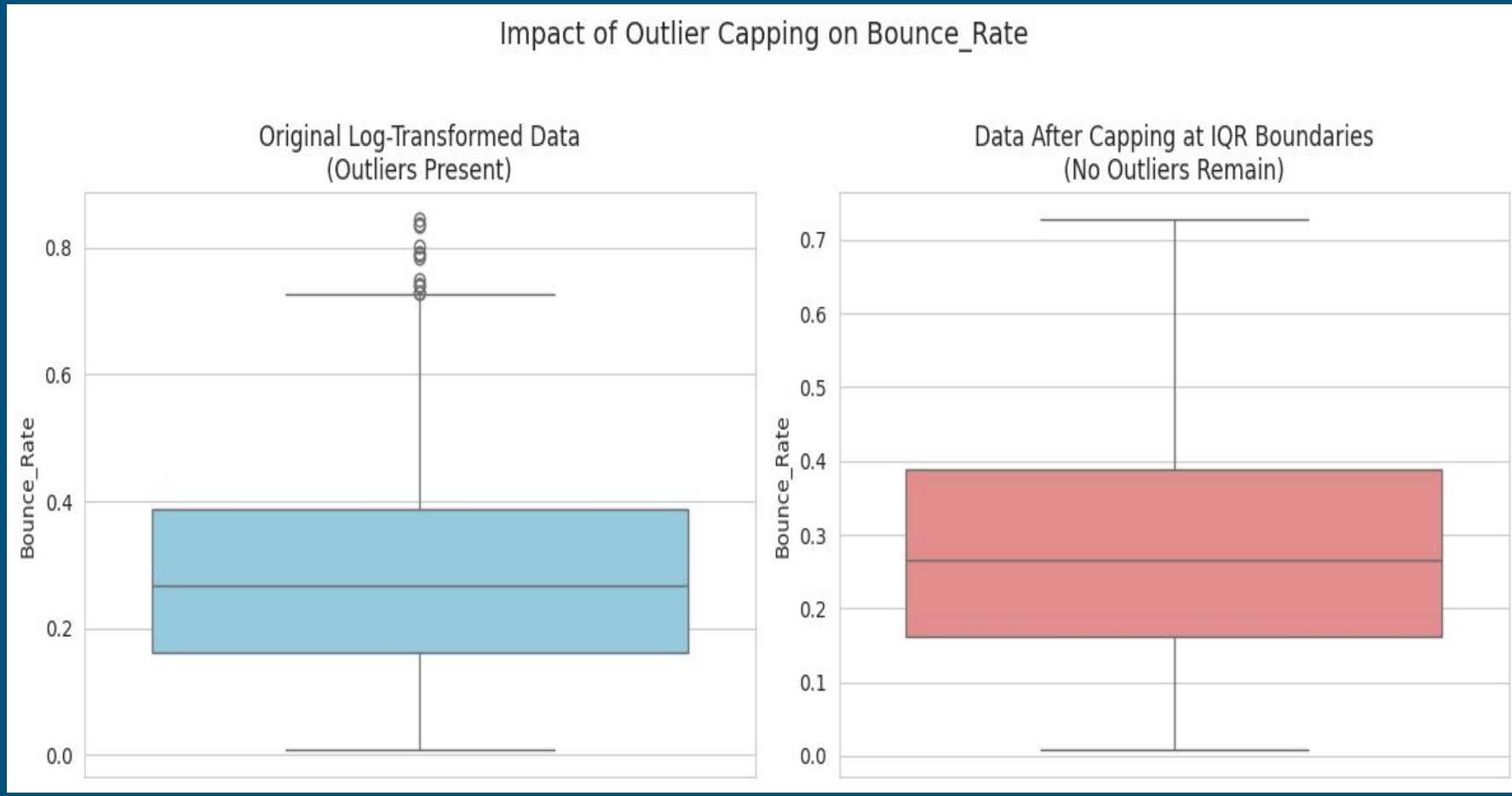


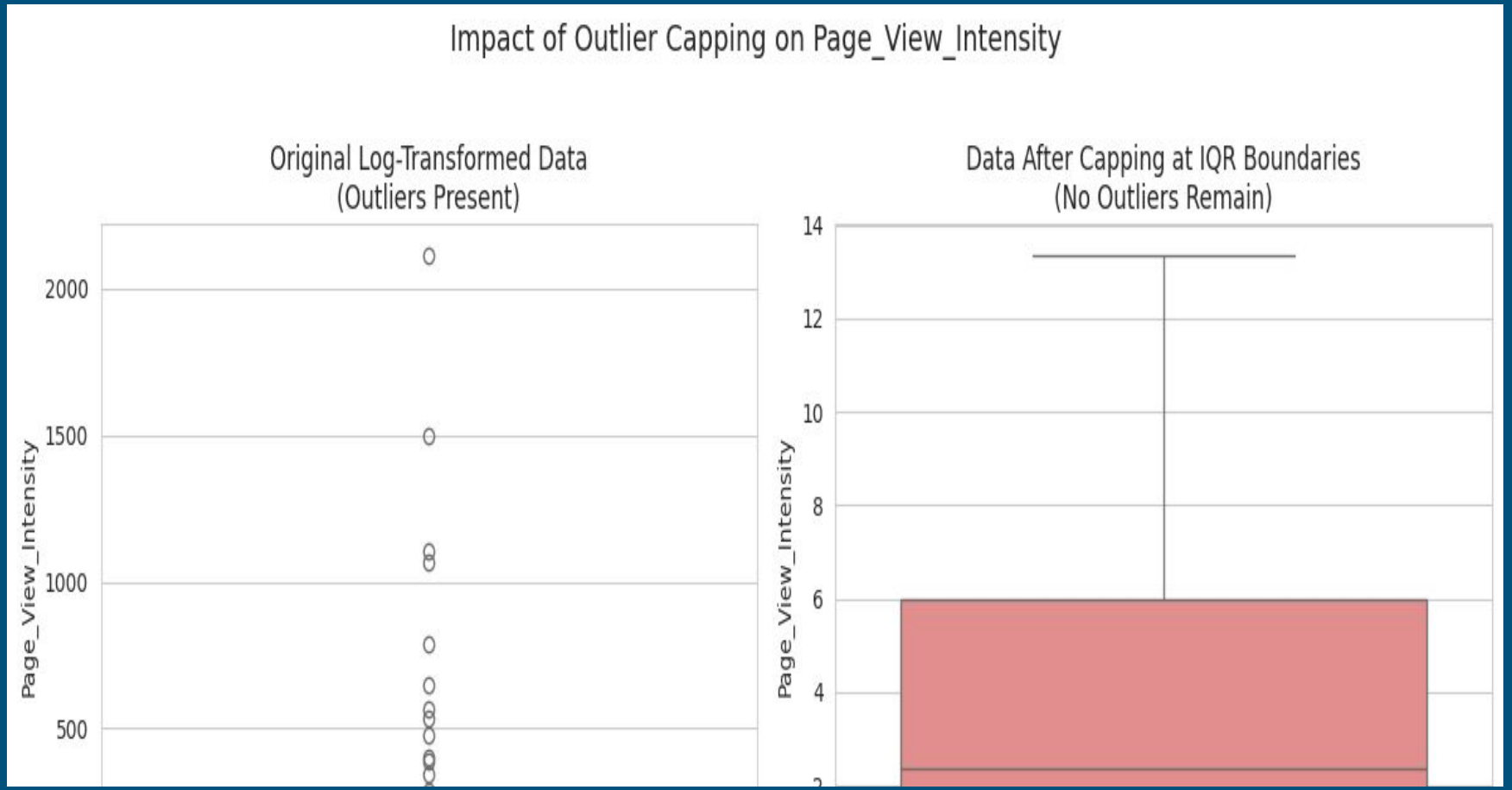



# Data Stabilization

**Outlier Management Via Capping**

- **Issue:** Outliers persisted, mainly in the newly engineered ratio features.
- **Action: Outlier Capping (Winsorizing)** was applied using the IQR rule.
- **Result:** Extreme values were limited, retaining all 1,986 sessions and ensuring a statistically stable dataset without distortion.


Impact of Outlier Capping on Bounce_Rate
Original Log-Transformed Data
(Outliers Present)
Bounce_Rate
0.0
0.2
0.4
0.6
0.8
Data After Capping at IQR Boundaries
(No Outliers Remain)
Bounce_Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


Impact of Outlier Capping on Page_View_Intensity
Original Log-Transformed Data
(Outliers Present)
Page_View_Intensity
2000
1500
1000
500
Data After Capping at IQR Boundaries
(No Outliers Remain)
Page_View_Intensity
14
12
10
8
6
4

# Actionable Business Intelligence

**Top Recommendations:**

1. **Optimize for Bounce Rate Reduction**
   - Critical predictor of conversion failure
   - Improve landing page design and relevance
2. **Focus on Organic Traffic**
   - Most reliable, high-volume source with low bounce rates
   - Invest in SEO optimization
3. **Enhance Time on Page**
   - Design engaging, high-impact content for key pages (checkout/forms)
   - Reduce friction in conversion funnel
4. **Target High Page View Intensity Users**
   - Identify and nurture purposeful, efficient browsers
   - These users know what they want and convert effectively

# Summary & Next Steps

**Key Takeaways:**

- Dataset cleaned and transformed (1,986 sessions, 14 features)
- **1.8% non-converting sessions** characterized by high bounce rates and low engagement
- **Traffic source** is less important than **engagement behavior**
- Engineered features (Page View Intensity, Last Page Share) provide unique insights

**Next Steps:**

- Build predictive models (Logistic Regression, Random Forest, XGBoost)
- Deploy model to identify at-risk sessions in real-time
- A/B test interventions to reduce bounce rate

# References


- Parwal, H., Radha, R. and Anita, X., 2025. An Experimental Study on Website Optimization: Test Your Strategies. *Procedia Computer Science*, *259*, pp.660-669.
- Filippou, G., Georgiadis, A.G. and Jha, A.K., 2024. Establishing the link: Does web traffic from various marketing channels influence direct traffic source purchases?. *Marketing Letters*, *35*(1), pp.59-71.
- Pande, P.V., Tarbani, N.M. and Ingalkar, P.V., 2014. A study of web traffic analysis. *International Journal of Computer Science and Mobile Computing*, *3*(3), pp.900-907.
- Samaan, S.S., Korial, A.E., Sarra, R.R. and Humaidi, A.J., 2025. Multilingual Web Traffic Forecasting for Network Management Using Artificial Intelligence Techniques. *Results in Engineering*, p.105262.
- Pai, Chen-Sheng, Ta-Shun Cho, and Shieh-Liang Chen. "TRAFFIC ANALYSIS OF E-COMMERCE WEBSITES: EXPLORING THE MEDIATING EFFECT OF CONSUMER BEHAVIOR." *International Journal of Organizational Innovation* 17, no. 3 (2025).