EXPLORATORY DATA ANALYSIS

ENGAGEMENT TRENDS

PREPARED BY

AKRITI SHARMA
EMBADTA24014

Exploratory Data Analysis of Engagement Metrics: Understanding User Interaction and Trends

The goal of this report is to perform **Exploratory Data Analysis (EDA)** on the provided *EngagementMetrics.csv* dataset. This analysis will involve examining the data, identifying trends and patterns, testing relevant hypotheses, and extracting actionable insights. By exploring key engagement metrics such as sessions, user activity, event counts, and engagement rates across various channels, we aim to uncover the underlying factors driving user engagement and inform data-driven decisions to enhance performance and optimize engagement strategies.

Objective

The objective of this analysis is to perform an exploratory data analysis (EDA) on the provided engagement metrics dataset. The focus is to examine key metrics such as user engagement, session activity, event counts, and engagement rates across various channels. The analysis aims to uncover trends, detect outliers, and understand the distribution of engagement over time and across channels, providing valuable insights to inform decision-making and optimize user engagement strategies

- How different channels impact overall engagement rates and user interaction.
- The correlation between session length, user activity, and engagement levels.
- Identifying time-based patterns and their effect on user engagement.
- Understanding the relationship between event frequency per session and engagement outcomes.
- Exploring user behavior and identifying high-engagement user segments.
- Evaluating the impact of total session counts on engaged session rates.
- Identifying trends in engagement rates across different periods (e.g., daily, weekly, monthly).
- Understanding how the average time spent per session influences the number of events triggered.
- Investigating the role of 'EngagedSessionsPerUser' in driving platform activity.
- Assessing how session activity and event count correlate with overall engagement.

Data Understanding

The dataset contains 3,183 records across 10 columns. These columns capture engagement metrics such as:

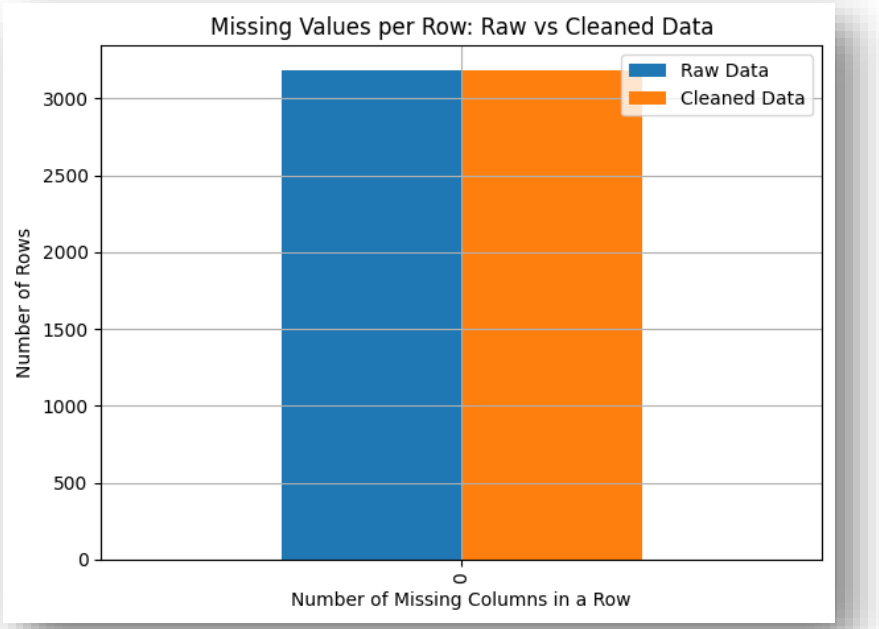
Attribute	Description	Data Type	Value Range
ChannelName	The channel through which the engagement occurred (e.g., Direct, Organic Social).	Categorical	Categories: Direct, Organic Social, etc.
DateTime	The timestamp (in the format YYYYMMDDHH) of when the engagement activity occurred.	DateTime	Range: 2024-04-16 23:00 to 2024-04-17 23:00
Users	The number of unique users involved in the engagement activity.	Numeric	Range: 0 to max unique users per record
Sessions	The total number of sessions recorded for the engagement activity.	Numeric	Range: 0 to max sessions per record

EngagedSessions	The number of sessions where users actively engaged with the content.	Numeric	Range: 0 to max engaged sessions per record
AvgEngagementTimePerSession	The average time users spend per session when engaged.	Numeric (Time)	Range: 0 to max average time per session
EngagedSessionsPerUser	The ratio of engaged sessions to the total number of users.	Numeric (Ratio)	Range: 0 to 1
EventsPerSession	The number of events triggered per session (e.g., clicks, views, interactions).	Numeric	Range: 0 to max events per session
EngagementRate	The ratio of engaged sessions to total sessions, expressed as a percentage.	Numeric (%)	Range: 0% to 100%
EventCount	The total number of events triggered across all sessions.	Numeric	Range: 0 to total event count in dataset

Data Cleaning and Preparation

Identify and manage missing or null values in the dataset.

- Detect missing values across all columns.
- Impute missing values in numeric columns using mean, median, or other appropriate methods.
- Ensure the dataset is complete and ready for analysis.

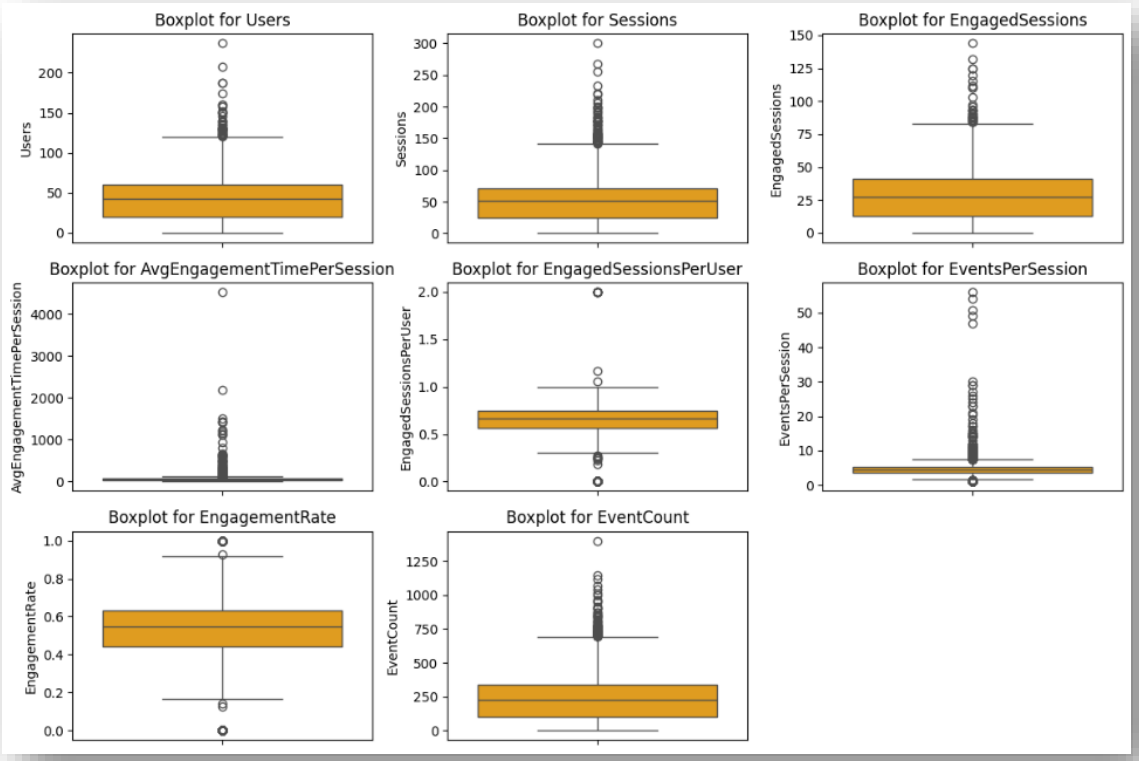


Outlier Detection and Treatment

Identify and handle outliers to ensure data consistency and accuracy. We will identify any extreme values or outliers in columns like AvgEngagementTimePerSession, EventCount, and EngagementRate, which may indicate errors or anomalies.

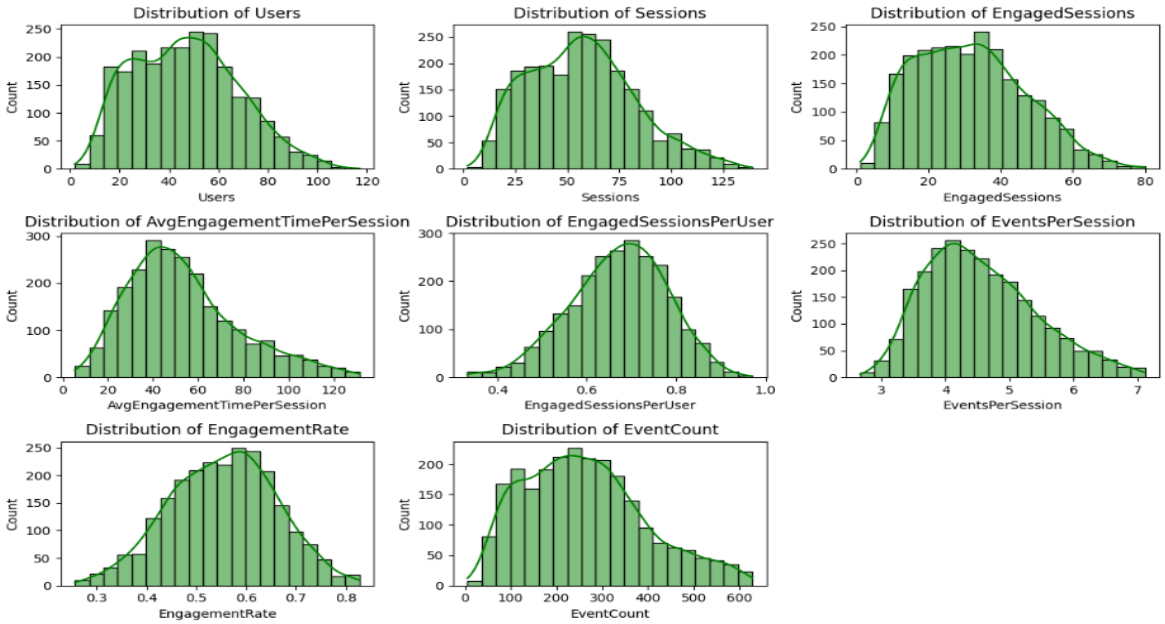
- Visualize data distributions using boxplots to identify outliers.
- Use the Interquartile Range (IQR) method to remove or adjust outliers in numeric columns.

The below boxplots display the distribution of engagement metrics like Users, Sessions, and EventCount. They highlight **outliers** in the data, which may require further investigation or removal.



Visualization of cleaned Data

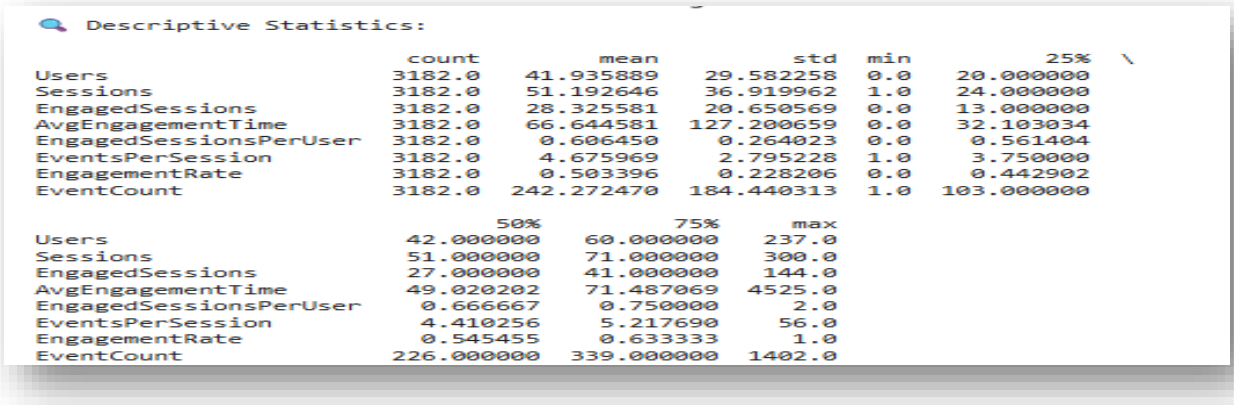
The histograms with KDE provide a clear view of the data distribution for metrics such as Sessions and EngagementRate, helping to identify trends, skewness, and patterns in user engagement.



Descriptive Statistics and Data Distribution

Generate and interpret basic summary statistics for numeric columns.

- Use .describe() to obtain the mean, median, standard deviation, min, max, and quartiles.
- Evaluate central tendencies and data dispersion to understand data behavior.



Descriptive Statistics:

	count	mean	std	min	25%	\
Users	3182.0	41.935889	29.582258	0.0	20.000000	
Sessions	3182.0	51.192646	36.919962	1.0	24.000000	
EngagedSessions	3182.0	28.325581	20.650569	0.0	13.000000	
AvgEngagementTime	3182.0	66.644581	127.200659	0.0	32.103034	
EngagedSessionsPerUser	3182.0	0.606450	0.264023	0.0	0.561404	
EventsPerSession	3182.0	4.675969	2.795228	1.0	3.750000	
EngagementRate	3182.0	0.503396	0.228206	0.0	0.442902	
EventCount	3182.0	242.272470	184.440313	1.0	103.000000	

	50%	75%	max
Users	42.000000	60.000000	237.0
Sessions	51.000000	71.000000	300.0
EngagedSessions	27.000000	41.000000	144.0
AvgEngagementTime	49.020202	71.487069	4525.0
EngagedSessionsPerUser	0.666667	0.750000	2.0
EventsPerSession	4.410256	5.217690	56.0
EngagementRate	0.545455	0.633333	1.0
EventCount	226.000000	339.000000	1402.0

Below are the observation :-

- **Right-Skewed Distributions**: Most metrics like AvgEngagementTime and EventCount show a right-skewed distribution, indicating that while most sessions have moderate engagement, a few sessions exhibit very high values.
- **High Variability**: Engagement metrics such as AvgEngagementTime and EventCount have significant variability, suggesting inconsistent user behavior or differing content types.
- **Outliers:** Extreme values, particularly in AvgEngagementTime (up to 4525 minutes) and EventCount (up to 1402 events), indicate the presence of outliers that may require further review.
- **Moderate Engagement:** The average engagement rate is around 50%, highlighting that while half of the sessions show meaningful engagement, there's room for improvement.
- **User Engagement:** On average, users have fewer than one engaged session, indicating potential areas to increase user interaction.

Metric	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Users	3182	41.94	29.58	0	20	42	60	237
Sessions	3182	51.19	36.92	1	24	51	71	300
Engaged Sessions	3182	28.33	20.65	0	13	27	41	144
Avg. Engagement Time	3182	66.64	127.2	0	32.1	49.02	71.49	4525
Engaged Sessions/User	3182	0.606	0.264	0	0.561	0.667	0.75	2
Events per Session	3182	4.68	2.8	1	3.75	4.41	5.22	56
Engagement Rate	3182	0.503	0.228	0	0.443	0.545	0.633	1
Event Count	3182	242.27	184.44	1	103	226	339	1402

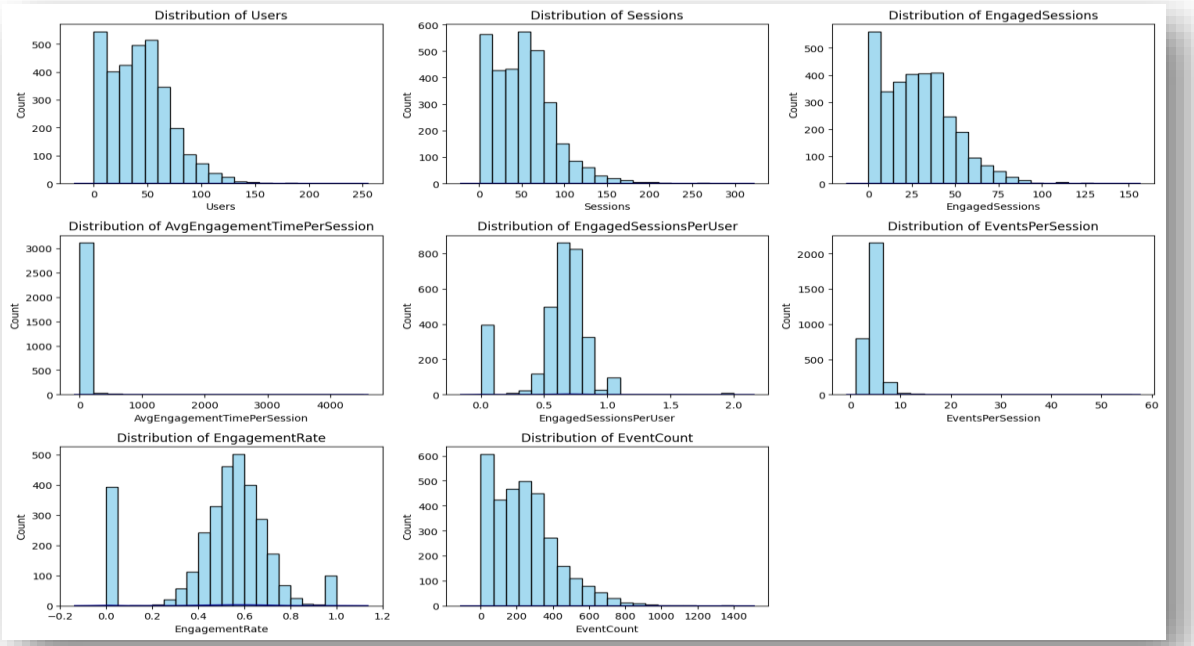
Visualizing the Data

The next step in EDA is visualizing the data. Graphical representations help uncover trends, relationships, and potential issues that may not be evident in summary statistics. We will use various plots to explore the dataset further.

Data Distribution Analysis:

Visualize the distribution of key metrics (e.g., Users, Sessions, EngagementRate) using histograms and KDE (Kernel Density Estimate).

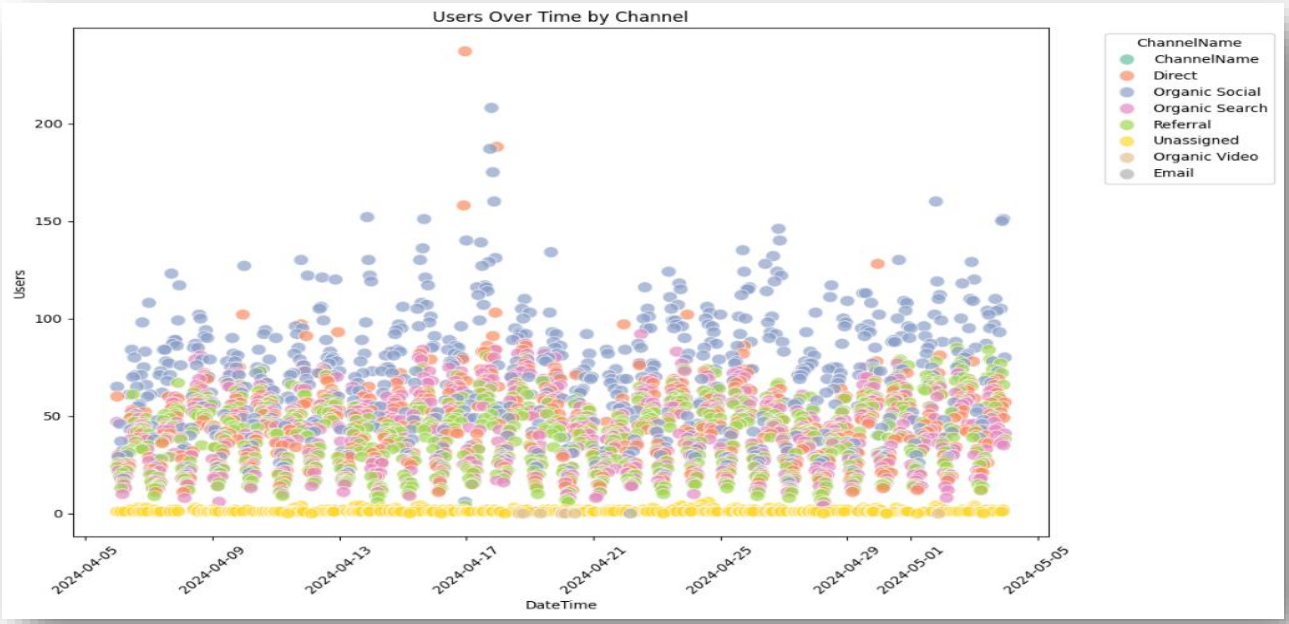
- **Right-skewed distributions**: Most metrics (e.g., Sessions, Users, EventCount) are right-skewed, indicating a small number of sessions have much higher activity levels than the majority.
- **Outliers and Anomalies**: Some metrics, like AvgEngagementTimePerSession and EventCount, show extreme values, suggesting that these outliers should be investigated further.
- **Engagement Patterns**: While many sessions are relatively short or have low engagement, there are some with exceptionally high engagement, pointing to potentially popular content or high user involvement.



Plot 1 : Users Over Time by Channel

The scatter plot now represents the number of users (Users) over time (user activity) and different colors are used for different channels (such as Direct, Organic Social, etc.).

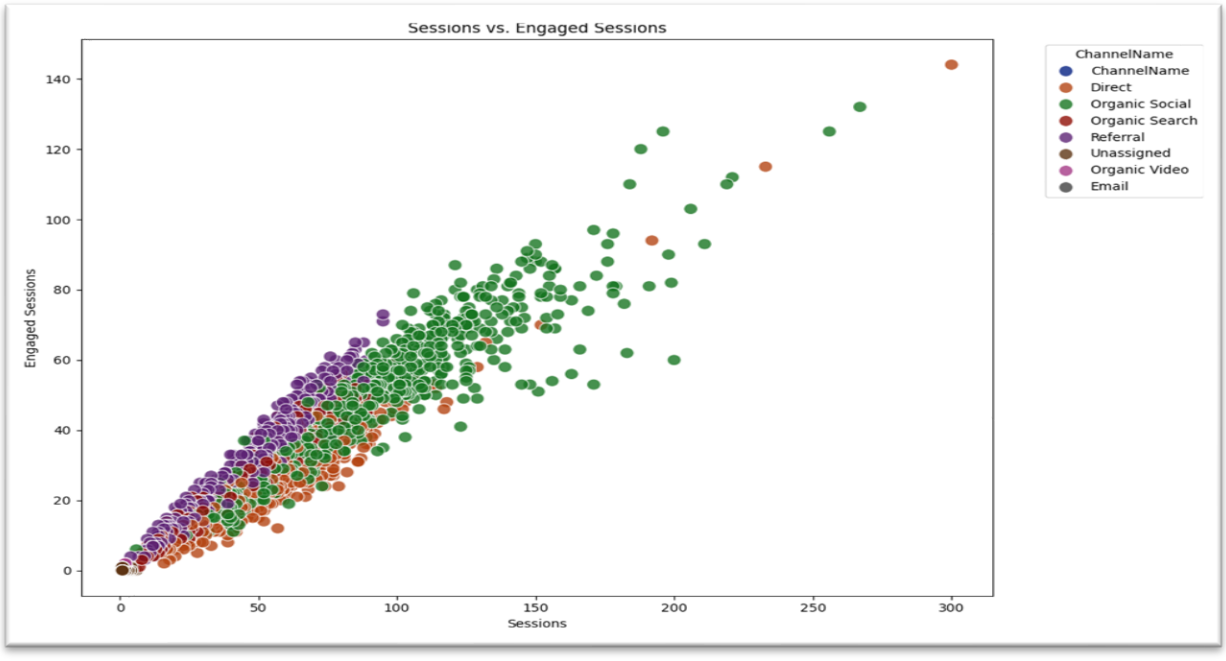
- Peaks and dips will highlight periods of high or low user engagement for specific channels.
- This visualization will help you track trends, such as which channels bring in more users and how user activity changes over time.



Plot 2 : Sessions vs. Engaged Sessions

This scatter plot visualizes the relationship between total sessions and engaged sessions across different marketing channels.

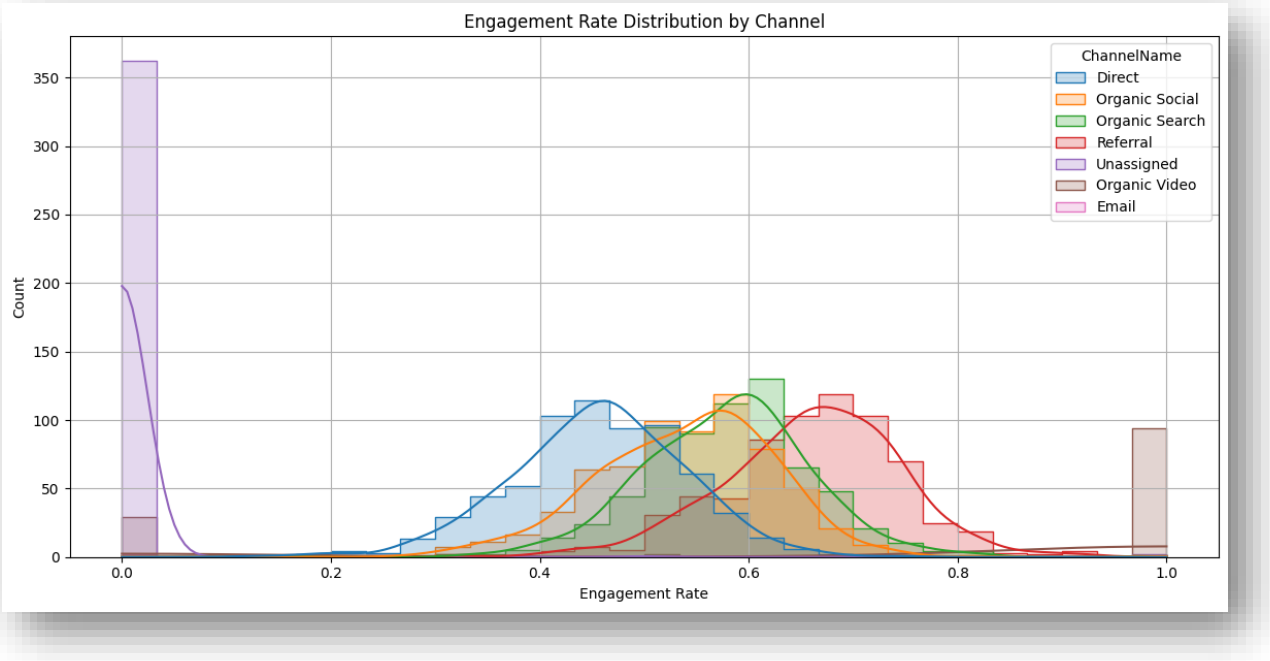
- Positive Relationship: The clear upward trend between total sessions and engaged sessions indicates that increased traffic generally leads to higher engagement across all channels.
- Channel Comparison: This plot allows you to visually assess which channels are performing better in terms of engagement. Channels like Organic Social and Organic Search seem to be more efficient at converting sessions into engaged sessions



Plot 3 : Sessions vs. Engaged Sessions

This chart shows how engagement rates are distributed across various marketing channels. Here's how to read it:

- Direct and Organic Search have a bell-shaped curve centered around moderate engagement rates (0.4–0.6) indicating consistency.
- Referral tends to show slightly higher engagement rates (peaking around 0.7) possibly more targeted traffic.
- Email has tight clustering near high engagement (closer to 1.0) suggesting a strong, loyal audience segment.
- Unassigned has a huge spike at 0, indicating a data capture or tagging issue should be reviewed.

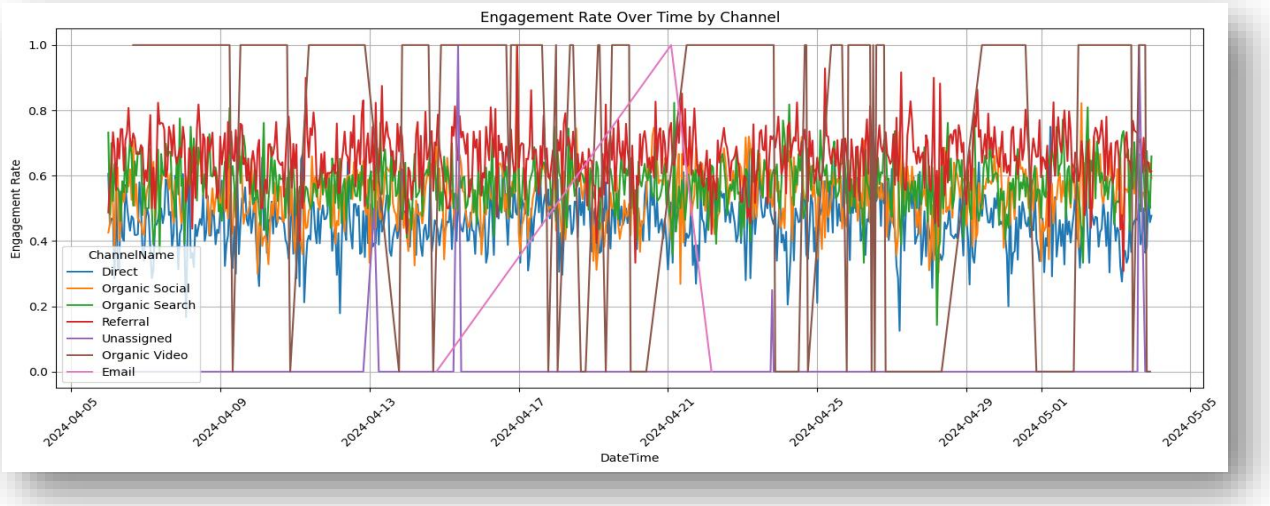


Plot 4 : Engagement Rate Over Time by Channel

This line chart shows how the engagement rate has varied hour-by-hour/day-by-day across different digital marketing channels.

- Referral consistently shows the highest engagement rate (~0.7), indicating strong user intent.
- Direct and Organic Social have stable but moderate engagement (~0.4–0.5).
- Unassigned and Organic Video show erratic spikes (0 or 1), suggesting tagging or data quality issues.
- Email shows a short peak, likely from a time-bound campaign with strong initial performance.

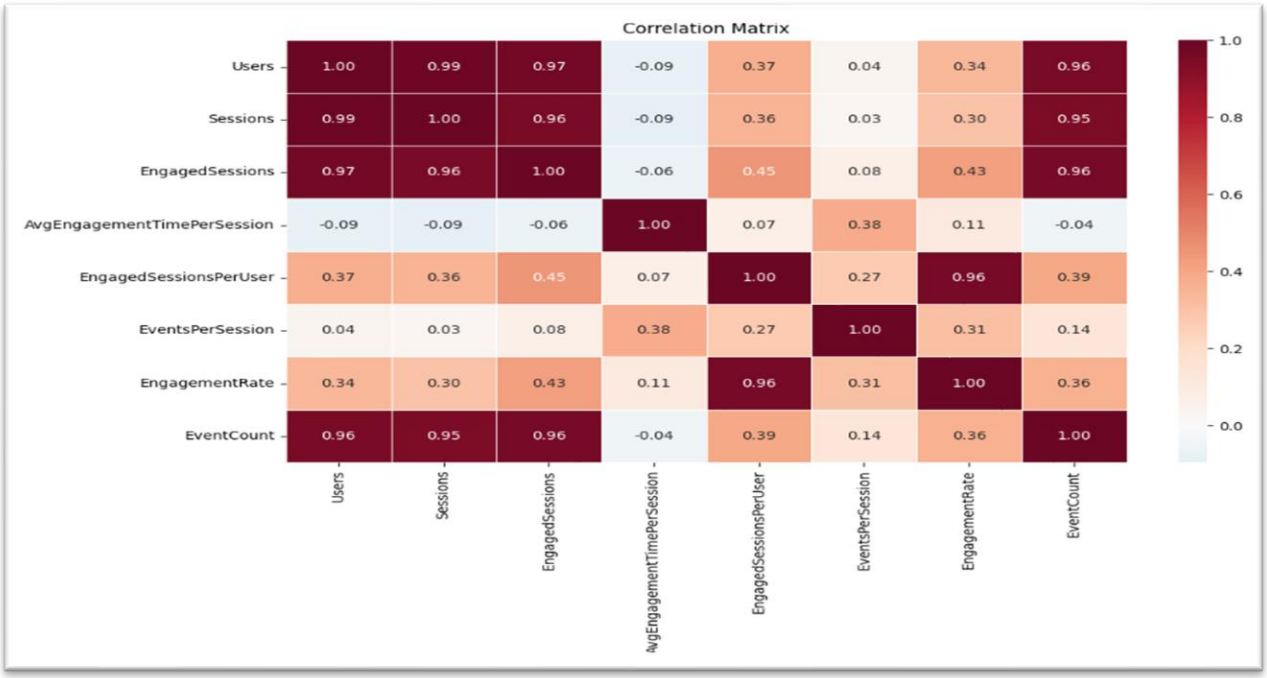
- Overall, Referral stands out as the most reliable high-engagement channel.



Plot 5 : Correlation Matrix

This heatmap shows the strength and direction of relationships between key engagement metrics. Dark cells indicate strong positive relationships

- Users, Sessions, Engaged Sessions, Event Count are highly correlated (~0.95–0.99) These metrics move together more users generally means more sessions and events.
- Engagement Rate has a very strong correlation (0.96) with Engaged Sessions per User. Engagement Rate is largely influenced by how many sessions per user are engaged.
- Avg Engagement Time shows low or negative correlation with most metrics Time spent doesn't always translate into more sessions or higher engagement suggesting quality is not equal to duration.
- Events per Session is weakly correlated with other metrics.Session activity varies widely across users/channels needs deeper segmentation.



Insights from Exploratory Data Analysis (EDA)

- Referral and Organic Social consistently show high engagement rates and rich session interactions, making them ideal channels for user retention and deep content consumption.
- Direct and Organic Search generate the highest number of users and sessions, but with moderate engagement, indicating that visitors are not staying long or interacting deeply.
- Organic Video has very high engagement per session, even with low traffic volume showing strong potential among a niche audience that resonates with visual content.
- Email channel shows brief spikes in engagement, suggesting success in short, time-bound campaigns but may need broader or more consistent execution.
- Unassigned traffic shows unusual behavior with extreme highs or zeros in engagement rate, which likely stems from missing or misconfigured tracking parameters.
- Engaged Sessions per User has the strongest correlation with Engagement Rate (0.96), making it a key KPI for measuring content effectiveness.
- Average Engagement Time has low correlation with engagement rate, proving that longer time doesn't always mean better interaction.
- Users, Sessions, and Event Count are tightly linked indicating that volume-based metrics rise together, but don't guarantee quality engagement.
- Events per Session varies significantly across channels and shows weak correlation with other metrics highlighting potential for session design improvements.
- Data highlights a clear distinction: Channels that bring traffic are not always the ones that drive engagement. Optimizing both is key for ROI.

Recommendations

- Scale Referral and Organic Social by investing in influencer content, personalization, and social SEO strategies.
- Fix tracking issues in Unassigned traffic through proper UTM tagging and analytics configuration.
- Improve Direct and Organic Search by optimizing landing pages, improving targeting, and refining content alignment.
- Strengthen Email campaigns with better audience segmentation, more engaging content, and consistent frequency.
- Expand Organic Video content to capitalize on its high engagement value among focused user groups.
- Use Engaged Sessions per User as a key KPI for measuring content and channel effectiveness.